

Recursive Sparse Recovery in Large but Structured Noise – Part 2

Chenlu Qiu and Namrata Vaswani

Abstract

We study the problem of recursively recovering a time sequence of sparse vectors, S_t , from measurements $M_t := S_t + L_t$ that are corrupted by structured noise L_t which is dense and can have large magnitude. The structure that we require is that L_t should lie in a low dimensional subspace that is either fixed or changes “slowly enough”; and the eigenvalues of its covariance matrix are “clustered”. We do not assume any model on the sequence of sparse vectors. Their support sets and their nonzero element values may be either independent or correlated over time (usually in many applications they are correlated). The only thing required is that there be *some* support change every so often. We introduce a novel solution approach called Recursive Projected Compressive Sensing with cluster-PCA (ReProCS-cPCA) that addresses some of the limitations of earlier work. Under mild assumptions, we show that, with high probability, ReProCS-cPCA can exactly recover the support set of S_t at all times; and the reconstruction errors of both S_t and L_t are upper bounded by a time-invariant and small value.

Keywords: robust PCA, sparse and low-rank matrix recovery, sparse recovery, compressive sensing

I. INTRODUCTION

In this work, we study the problem of recursively recovering a time sequence of sparse vectors, S_t , from measurements $M_t := S_t + L_t$ that are corrupted by structured noise L_t which is dense and can have large magnitude. The structure that we require is that L_t should lie in a low dimensional subspace that is either fixed or changes “slowly enough” as discussed in Sec II-B; and the eigenvalues of its covariance matrix are “clustered” as explained in Sec II-D. As a by-product, at certain times, the basis vectors for the subspace in which the most recent several L_t ’s lies is also recovered. Thus, at these times, we also solve the recursive robust principal components’ analysis (PCA) problem. For the recursive robust PCA problem, L_t is the signal of interest while S_t can be interpreted as the outlier (large but sparse noise).

A key application where the above problem occurs is in video analysis where the goal is to separate a slowly changing background from moving foreground objects [1], [2]. If one stacks each image frame as a column vector, the background is well modeled as lying in a low dimensional subspace that may gradually change over time, while the moving foreground objects constitute the sparse vectors [3], [2] which change in a correlated fashion over time. Another key application is online detection of brain activation patterns from functional MRI (fMRI) sequences. In this case, the “active” region of the brain is the correlated sparse vector.

A. Related Work

Many of the older works on sparse recovery with structured noise study the case of sparse recovery from large but sparse noise (outliers), e.g., [3], [4], [5]. However, here we are interested in sparse recovery in large but low dimensional noise. On the other hand, most older works on robust PCA cannot recover the outlier (S_t) when its nonzero entries have magnitude much smaller than that of the low dimensional part (L_t) [6], [1], [7]. The main goal of this work is to study sparse recovery and hence we do not discuss these older works here. Some recent works on robust PCA such as [8], [9] assume that an entire measurement vector M_t is either an inlier (S_t is a zero vector) or an outlier (all entries of S_t can be nonzero), and a certain number of M_t ’s are inliers. These works also cannot be used when all S_t ’s are nonzero but sparse.

In a series of recent works [2], [10], a new and elegant solution, which is referred to as Principal Components' Pursuit (PCP) in [2], has been proposed. It redefines batch robust PCA as a problem of separating a low rank matrix, $\mathcal{L}_t := [L_1, \dots, L_t]$, from a sparse matrix, $\mathcal{S}_t := [S_1, \dots, S_t]$, using the measurement matrix, $\mathcal{M}_t := [M_1, \dots, M_t] = \mathcal{L}_t + \mathcal{S}_t$. Thus these works can be interpreted as batch solutions to sparse recovery in large but low dimensional noise. Other recent works that also study batch algorithms for recovering a sparse \mathcal{S}_t and a low-rank \mathcal{L}_t from $\mathcal{M}_t := \mathcal{L}_t + \mathcal{S}_t$ or from undersampled measurements include [11], [12], [13], [14], [15], [16], [17], [18], [19], [20].

It was shown in [2] that, with high probability (w.h.p.), one can recover \mathcal{L}_t and \mathcal{S}_t exactly by solving

$$\min_{\mathcal{L}, \mathcal{S}} \|\mathcal{L}\|_* + \lambda \|\mathcal{S}\|_{1, \text{vec}} \quad \text{subject to} \quad \mathcal{L} + \mathcal{S} = \mathcal{M}_t \quad (1)$$

provided that (a) \mathcal{L}_t is dense (its left and right singular vectors satisfy certain conditions); (b) any element of the matrix \mathcal{S}_t is nonzero w.p. ϱ , and zero w.p. $1 - \varrho$, independent of all others (in particular, this means that the support sets of the different \mathcal{S}_t 's are independent over time); and (c) the rank of \mathcal{L}_t and the support size of \mathcal{S}_t are small enough. Here $\|B\|_*$ is the nuclear norm of B (sum of singular values of B) while $\|B\|_{1, \text{vec}}$ is the ℓ_1 norm of B seen as a long vector. In most applications, it is fair to assume that the low dimensional part, L_t (background in case of video) is dense. However, the assumption that the support of the sparse part (foreground in case of video) is independent over time is often not valid. Foreground objects typically move in a correlated fashion, and may even not move for a few frames. This results in \mathcal{S}_t being sparse and low rank.

The question then is, what can we do if \mathcal{L}_t is low rank and dense, but \mathcal{S}_t is sparse and may also be low rank? In this case, without any extra information, in general, it is not possible to separate \mathcal{S}_t and \mathcal{L}_t . In [21], we introduced the Recursive Projected Compressive Sensing (ReProCS) algorithm that provided one possible solution to this problem by using the extra piece of information that an initial short sequence of L_t 's, or L_t 's in small noise, is available (which can be used to get an accurate estimate of the subspace in which the initial L_t 's lie) and assuming slow subspace change (as explained in Sec. II-B). The key idea of ReProCS is as follows. At time t , assume that a $n \times r$ matrix with orthonormal columns, $\hat{P}_{(t-1)}$, is available with $\text{span}(\hat{P}_{(t-1)}) \approx \text{span}(\mathcal{L}_{t-1})$. We project M_t perpendicular to $\text{span}(\hat{P}_{(t-1)})$. Because of slow subspace change, this cancels out most of the contribution of L_t . Recovering \mathcal{S}_t from the projected measurements then becomes a classical sparse recovery / compressive sensing (CS) problem in small noise [22]. Under a denseness assumption on $\text{span}(\mathcal{L}_{t-1})$, one can show that \mathcal{S}_t can be accurately recovered via ℓ_1 minimization. Thus, $L_t = M_t - \mathcal{S}_t$ can also be recovered accurately. We use the estimates of L_t in a projection-PCA based subspace estimation algorithm to update $\hat{P}_{(t)}$.

ReProCS is designed under the assumption that the subspace in which the most recent several L_t 's lie can only grow over time. It assumes a model in which at every subspace change time, t_j , some new directions get added to this subspace. After every subspace change, it uses projection-PCA to estimate the newly added subspace. As a result the rank of $\hat{P}_{(t)}$ keeps increasing with every subspace change. Therefore, the number of effective measurements available for the CS step, $(n - \text{rank}(\hat{P}_{(t-1)}))$, keeps reducing. To keep this number large enough at all times, ReProCS needs to assume a bound on the total number of subspace changes, J .

B. Our Contributions and More Related Work

In practice, usually, the dimension of the subspace in which the most recent several L_t 's lie typically remains roughly constant. A simple way to model this is to assume that at every change time, t_j , some new directions can get added and some existing directions can get deleted from this subspace and to assume an upper bound on the difference between the total number of added and deleted directions (the earlier model in [21] is a special case of this). ReProCS still applies for this more general model as discussed in the extensions section of [21]. However, because it never deletes directions, the rank of $\hat{P}_{(t)}$ still keeps increasing with every subspace change time and so it still requires a bound on J .

In this work, we address the above limitation by introducing a novel approach called *cluster-PCA* that re-estimates the current subspace after the newly added directions have been accurately estimated. This re-estimation step ensures that the deleted directions have been "removed" from the new $\hat{P}_{(t)}$. We refer to the resulting algorithm as *ReProCS-cPCA*. The design and analysis of cluster-PCA and ReProCS-cPCA is the focus of the current paper. We will see that ReProCS-cPCA does not

need a bound on J as long as the delay between subspace change times increases in proportion to $\log J$. An extra assumption that is needed though is that the eigenvalues of the covariance matrix of L_t are sufficiently clustered at certain times as explained in Sec II-D. As discussed in Sec IV-B, this is a practically valid assumption.

Under the clustering assumption and some other mild assumptions, we show that, w.h.p, at all times, ReProCS-cPCA can exactly recover the support of S_t , and the reconstruction errors of both S_t and L_t are upper bounded by a time invariant and small value. Moreover, we show that the subspace recovery error decays roughly exponentially with every projection-PCA step. The proof techniques developed in this work are very different from those used to obtain performance guarantees in recent batch robust PCA works such as [2], [10], [23], [8], [9], [11], [12], [19], [17], [16], [18], [20]. As explained earlier, [8], [9] also study a different problem. Our proof utilizes sparse recovery results [22]; results from matrix perturbation theory (sin θ theorem [24] and Weyl's theorem [25]) and the matrix Hoeffding inequality [26].

Our result for ReProCS-cPCA (and also that for ReProCS from [21]) does not assume any model on the sparse vectors', S_t 's. In particular, it allows the support sets of the S_t 's to be either independent, e.g. generated via the model of [2] (resulting in S_t being full rank w.h.p.), or correlated over time (can result in S_t being low rank). As explained in Sec IV-B, the only thing that is required is that there be *some* support changes every so often. We should point out that some of the other works that study the batch problem, e.g. [2], [16], also allow S_t to be low rank.

A key difference of our work compared with most existing work analyzing finite sample PCA, e.g. [27], and references therein, is that in these works, the noise/error in the observed data is independent of the true (noise-free) data. However, in our case, because of how \hat{L}_t is computed, the error $e_t = L_t - \hat{L}_t$ is correlated with L_t . As a result the tools developed in these earlier works cannot be used for our problem. This is the main reason we need to develop and analyze projection-PCA based approaches for both subspace addition and deletion.

In earlier conference papers [28], [29], we first introduced the ReProCS idea. However, these used an algorithm motivated by recursive PCA [30] for updating the subspace estimates on-the-fly. As explained in Sec III and also in [21, Appendix F], it is not clear how to obtain performance guarantees for recursive PCA (which is a fast algorithm for PCA) for our problem. Another online algorithm that addresses a problem similar to ours is given in [31]. This also does not obtain guarantees.

The ReProCS-cPCA approach is related to that of [32], [33], [34] in that all of these first try to nullify the low dimensional signal by projecting the measurement vector into a subspace perpendicular to that of the low dimensional signal, and then solve for the sparse "error" vector. However, the big difference is that in all of these works the basis for the subspace of the low dimensional signal is *perfectly known*. We study *the case where the subspace is not known and can change over time*.

C. Paper Organization

We give the notation next followed by a review of results from existing work that we will need. The problem definition and the three key assumptions that are needed are explained in Sec II. We develop the ReProCS-cPCA algorithm in Sec III. We give its performance guarantees (Theorem 4.1) in Sec IV. Here we also provide a discussion of the result and the assumptions it makes. We define the quantities needed for the proof and give the proof outline in Sec V. The proof of Theorem 4.1 is given in Sec VI. The key lemmas needed for it are given and proved in Sec VII. In Sec VIII, we show numerical experiments demonstrating Theorem 4.1, as well as comparisons with ReProCS and PCP. Conclusions are given in Sec IX.

D. Notation

For a set $T \subseteq \{1, 2, \dots, n\}$, we use $|T|$ to denote its cardinality, i.e., the number of elements in T . We use T^c to denote its complement w.r.t. $\{1, 2, \dots, n\}$, i.e. $T^c := \{i \in \{1, 2, \dots, n\} : i \notin T\}$. The notations $T_1 \subseteq T_2$ and $T_2 \supseteq T_1$ both mean that T_1 is a subset of T_2 .

We use the notation $[t_1, t_2]$ to denote an interval which contains t_1 and t_2 , as well as all integers between them, i.e. $[t_1, t_2] := \{t_1, t_1 + 1, \dots, t_2\}$. The notation $[L_t; t \in [t_1, t_2]]$ is used to denote the matrix $[L_{t_1}, L_{t_1+1}, \dots, L_{t_2}]$.

For a vector v , v_i denotes the i th entry of v and v_T denotes a vector consisting of the entries of v indexed by T . We use $\|v\|_p$ to denote the ℓ_p norm of v . The support of v , $\text{supp}(v)$, is the set of indices at which v is nonzero, $\text{supp}(v) := \{i : v_i \neq 0\}$. We say that v is s -sparse if $|\text{supp}(v)| \leq s$.

For a tall matrix P , $\text{span}(P)$ denotes the subspace spanned by the column vectors of P .

For a matrix B , B' denotes its transpose, and B^\dagger denotes its pseudo-inverse. For a matrix with linearly independent columns, $B^\dagger = (B'B)^{-1}B'$. We use $\|B\|_2 := \max_{x \neq 0} \|Bx\|_2 / \|x\|_2$ to denote the induced 2-norm of the matrix. Also, $\|B\|_*$ is the nuclear norm and $\|B\|_{\max}$ denotes the maximum over the absolute values of all its entries. We let $\sigma_i(B)$ denote the i th largest singular value of B . For a Hermitian matrix, B , we use the notation $B \stackrel{EVD}{=} U\Lambda U'$ to denote the eigenvalue decomposition (EVD) of B . Here U is an orthonormal matrix and Λ is a diagonal matrix with entries arranged in non-increasing order. Also, we use $\lambda_i(B)$ to denote the i th largest eigenvalue of a Hermitian matrix B and we use $\lambda_{\max}(B)$ and $\lambda_{\min}(B)$ denote its maximum and minimum eigenvalues. If B is Hermitian positive semi-definite (p.s.d.), then $\lambda_i(B) = \sigma_i(B)$. For Hermitian matrices B_1 and B_2 , the notation $B_1 \preceq B_2$ means that $B_2 - B_1$ is p.s.d. Similarly, $B_1 \succeq B_2$ means that $B_1 - B_2$ is p.s.d.

For a Hermitian matrix B , we have $\|B\|_2 = \sqrt{\max(\lambda_{\max}^2(B), \lambda_{\min}^2(B))}$. Thus, for a $b \geq 0$, $\|B\|_2 \leq b$ implies that $-b \leq \lambda_{\min}(B) \leq \lambda_{\max}(B) \leq b$. If B is a Hermitian p.s.d. matrix, then $\|B\|_2 = \lambda_{\max}(B)$.

The notation $[\cdot]$ denotes an empty matrix. We use I to denote an identity matrix. For an $m \times n$ matrix B and an index set $T \subseteq \{1, 2, \dots, n\}$, B_T is the sub-matrix of B containing columns with indices in the set T . Notice that $B_T = BI_T$. We use $B \setminus B_T$ to denote B_{T^c} (here $T^c := \{i \in \{1, 2, \dots, n\} : i \notin T\}$). Given another matrix B_2 of size $m \times n_2$, $[B \ B_2]$ constructs a new matrix by concatenating matrices B and B_2 in horizontal direction. Thus, $[(B \setminus B_T) \ B_2] = [B_{T^c} \ B_2]$. For any matrix B and sets T_1, T_2 , $(B)_{T_1, T_2}$ denotes the sub-matrix containing the rows with indices in T_1 and columns with indices in T_2 .

Definition 1.1: We refer to a tall matrix P as a *basis matrix* if it satisfies $P'P = I$.

Definition 1.2: The *s-restricted isometry constant (RIC)* [32], δ_s , for an $n \times m$ matrix Ψ is the smallest real number satisfying $(1 - \delta_s)\|x\|_2^2 \leq \|\Psi_T x\|_2^2 \leq (1 + \delta_s)\|x\|_2^2$ for all sets $T \subseteq \{1, 2, \dots, n\}$ with $|T| \leq s$ and all real vectors x of length $|T|$.

It is easy to see that $\max_{T: |T| \leq s} \|(\Psi_T' \Psi_T)^{-1}\|_2 \leq \frac{1}{1 - \delta_s(\Psi)}$ [32].

Definition 1.3: Let X and Z be two random variables (r.v.) and let \mathcal{B} be a set of values that Z can take.

- 1) We use \mathcal{B}^e to denote the *event* $Z \in \mathcal{B}$, i.e. $\mathcal{B}^e := \{Z \in \mathcal{B}\}$.
- 2) The probability of event \mathcal{B}^e can be expressed as [35],

$$\mathbf{P}(\mathcal{B}^e) := \mathbf{E}[\mathbb{I}_{\mathcal{B}}(Z)].$$

where

$$\mathbb{I}_{\mathcal{B}}(Z) := \begin{cases} 1 & \text{if } Z \in \mathcal{B} \\ 0 & \text{otherwise} \end{cases}$$

is an indicator function of Z on the set \mathcal{B} and $\mathbf{E}[\mathbb{I}_{\mathcal{B}}(Z)]$ is the expectation of $\mathbb{I}_{\mathcal{B}}(Z)$.

- 3) Define $\mathbf{P}(\mathcal{B}^e|X) := \mathbf{E}[\mathbb{I}_{\mathcal{B}}(Z)|X]$ where $\mathbf{E}[\mathbb{I}_{\mathcal{B}}(Z)|X]$ is the conditional expectation of $\mathbb{I}_{\mathcal{B}}(Z)$ given X .

Finally, RHS refers to the right hand side of an equation or inequality; w.p. means “with probability”; and w.h.p. means “with high probability”.

E. Preliminaries

In this section we state certain results from literature, or certain lemmas which follow easily using these results, that will be used in proving our main result.

- 1) *Simple probability facts and matrix Hoeffding inequalities:* The following result follows directly from Definition 1.3.

Lemma 1.4: Suppose that \mathcal{B} is the set of values that the r.v.s X, Y can take. Suppose that \mathcal{C} is a set of values that the r.v. X can take. For a $0 \leq p \leq 1$, if $\mathbf{P}(\mathcal{B}^e|X) \geq p$ for all $X \in \mathcal{C}$, then $\mathbf{P}(\mathcal{B}^e|\mathcal{C}^e) \geq p$ as long as $\mathbf{P}(\mathcal{C}^e) > 0$.

Proof: This is the same as [21, Lemma 11].

The following lemma is an easy consequence of the chain rule of probability applied to a contracting sequence of events.

Lemma 1.5: For a sequence of events $E_0^e, E_1^e, \dots, E_m^e$ that satisfy $E_0^e \supseteq E_1^e \supseteq E_2^e \dots \supseteq E_m^e$, the following holds

$$\mathbf{P}(E_m^e | E_0^e) = \prod_{k=1}^m \mathbf{P}(E_k^e | E_{k-1}^e).$$

Proof: $\mathbf{P}(E_m^e | E_0^e) = \mathbf{P}(E_m^e, E_{m-1}^e, \dots, E_0^e | E_0^e) = \prod_{k=1}^m \mathbf{P}(E_k^e | E_{k-1}^e, E_{k-2}^e, \dots, E_0^e) = \prod_{k=1}^m \mathbf{P}(E_k^e | E_{k-1}^e)$. \blacksquare

The following two results are corollaries of the matrix Hoeffding inequality [26, Theorem 1.3] that were proved in [21]. In the rest of the paper we often refer to them as the *Hoeffding corollaries*.

Corollary 1.6 (Matrix Hoeffding conditioned on another random variable for a nonzero mean Hermitian matrix): Given an α -length sequence $\{Z_t\}$ of random Hermitian matrices of size $n \times n$, a r.v. X , and a set \mathcal{C} of values that X can take. Assume that, for all $X \in \mathcal{C}$, (i) Z_t 's are conditionally independent given X ; (ii) $\mathbf{P}(b_1 I \preceq Z_t \preceq b_2 I | X) = 1$ and (iii) $b_3 I \preceq \frac{1}{\alpha} \sum_t \mathbf{E}(Z_t | X) \preceq b_4 I$. Then for all $\epsilon > 0$,

$$\begin{aligned} \mathbf{P}(\lambda_{\max}(\frac{1}{\alpha} \sum_t Z_t) \leq b_4 + \epsilon | X) &\geq 1 - n \exp(-\frac{\alpha \epsilon^2}{8(b_2 - b_1)^2}) \text{ for all } X \in \mathcal{C} \\ \mathbf{P}(\lambda_{\min}(\frac{1}{\alpha} \sum_t Z_t) \geq b_3 - \epsilon | X) &\geq 1 - n \exp(-\frac{\alpha \epsilon^2}{8(b_2 - b_1)^2}) \text{ for all } X \in \mathcal{C} \end{aligned}$$

Proof: This is slight modification of [21, Corollary 13].

Corollary 1.7 (Matrix Hoeffding conditioned on another random variable for an arbitrary nonzero mean matrix): Given an α -length sequence $\{Z_t\}$ of random Hermitian matrices of size $n \times n$, a r.v. X , and a set \mathcal{C} of values that X can take. Assume that, for all $X \in \mathcal{C}$, (i) Z_t 's are conditionally independent given X ; (ii) $\mathbf{P}(\|Z_t\|_2 \leq b_1 | X) = 1$ and (iii) $\|\frac{1}{\alpha} \sum_t \mathbf{E}(Z_t | X)\|_2 \leq b_2$. Then, for all $\epsilon > 0$,

$$\mathbf{P}(\|\frac{1}{\alpha} \sum_t Z_t\|_2 \leq b_2 + \epsilon | X) \geq 1 - (n_1 + n_2) \exp(-\frac{\alpha \epsilon^2}{32b_1^2}) \text{ for all } X \in \mathcal{C}$$

Proof: This is slight modification of [21, Corollary 14].

2) *Linear algebra results:* Kahan and Davis's $\sin \theta$ theorem [24] studies the effect of a Hermitian perturbation, \mathcal{H} , on a Hermitian matrix, \mathcal{A} .

Theorem 1.8 ($\sin \theta$ theorem [24]): Given two Hermitian matrices \mathcal{A} and \mathcal{H} satisfying

$$\mathcal{A} = \begin{bmatrix} E & E_{\perp} \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & A_{\perp} \end{bmatrix} \begin{bmatrix} E' \\ E_{\perp}' \end{bmatrix}, \quad \mathcal{H} = \begin{bmatrix} E & E_{\perp} \end{bmatrix} \begin{bmatrix} H & B' \\ B & H_{\perp} \end{bmatrix} \begin{bmatrix} E' \\ E_{\perp}' \end{bmatrix} \quad (2)$$

where $\begin{bmatrix} E & E_{\perp} \end{bmatrix}$ is an orthonormal matrix. The two ways of representing $\mathcal{A} + \mathcal{H}$ are

$$\mathcal{A} + \mathcal{H} = \begin{bmatrix} E & E_{\perp} \end{bmatrix} \begin{bmatrix} A + H & B' \\ B & A_{\perp} + H_{\perp} \end{bmatrix} \begin{bmatrix} E' \\ E_{\perp}' \end{bmatrix} = \begin{bmatrix} F & F_{\perp} \end{bmatrix} \begin{bmatrix} \Lambda & 0 \\ 0 & \Lambda_{\perp} \end{bmatrix} \begin{bmatrix} F' \\ F_{\perp}' \end{bmatrix}$$

where $\begin{bmatrix} F & F_{\perp} \end{bmatrix}$ is another orthonormal matrix. Let $\mathcal{R} := (\mathcal{A} + \mathcal{H})E - \mathcal{A}E = \mathcal{H}E$. If $\lambda_{\min}(A) > \lambda_{\max}(\Lambda_{\perp})$, then

$$\|(I - FF')E\|_2 \leq \frac{\|\mathcal{R}\|_2}{\lambda_{\min}(A) - \lambda_{\max}(\Lambda_{\perp})}$$

Next we state the Weyl's theorem (Weyl's inequality for matrices) [25, page 181] and the Ostrowski's theorem [25, page 224].

Theorem 1.9 (Weyl [25]): Let \mathcal{A} and \mathcal{H} be two $n \times n$ Hermitian matrices. For each $i = 1, 2, \dots, n$ we have

$$\lambda_i(\mathcal{A}) + \lambda_{\min}(\mathcal{H}) \leq \lambda_i(\mathcal{A} + \mathcal{H}) \leq \lambda_i(\mathcal{A}) + \lambda_{\max}(\mathcal{H})$$

Theorem 1.10 (Ostrowski [25]): Let H and W be $n \times n$ matrices, with H Hermitian and W nonsingular. For each $i = 1, 2, \dots, n$, there exists a positive real number θ_i such that $\lambda_{\min}(WW') \leq \theta_i \leq \lambda_{\max}(WW')$ and $\lambda_i(WHW') = \theta_i \lambda_i(H)$. Therefore,

$$\lambda_{\min}(WHW') \geq \lambda_{\min}(WW') \lambda_{\min}(H)$$

The following lemma uses the $\sin \theta$ theorem and Weyl's theorem. It generalizes the idea of [21, Lemma 30].

Lemma 1.11: Suppose that two Hermitian matrices \mathcal{A} and \mathcal{H} can be decomposed as in (2) where $[E \ E_\perp]$ is an orthonormal matrix and A is a $c \times c$ matrix. Also, suppose that the EVD of $\mathcal{A} + \mathcal{H}$ is

$$\mathcal{A} + \mathcal{H} \stackrel{EVD}{=} \begin{bmatrix} F & F_\perp \end{bmatrix} \begin{bmatrix} \Lambda & 0 \\ 0 & \Lambda_\perp \end{bmatrix} \begin{bmatrix} F' \\ F'_\perp \end{bmatrix}$$

where Λ is a $c \times c$ diagonal matrix. If $\lambda_{\min}(A) > \lambda_{\max}(A_\perp) + \|\mathcal{H}\|_2$, then

$$\|(I - FF')E\|_2 \leq \frac{\|\mathcal{H}\|_2}{\lambda_{\min}(A) - \lambda_{\max}(A_\perp) - \|\mathcal{H}\|_2}$$

Proof: By definition of EVD, $[F \ F_\perp]$ is an orthonormal matrix. By the $\sin \theta$ theorem, if $\lambda_{\min}(A) > \lambda_{\max}(A_\perp)$, then $\|(I - FF')E\|_2 \leq \frac{\|\mathcal{R}\|_2}{\lambda_{\min}(A) - \lambda_{\max}(A_\perp)}$ where $\mathcal{R} := \mathcal{H}E$. Clearly, $\|\mathcal{R}\|_2 \leq \|\mathcal{H}\|_2$. Since $\lambda_{\min}(A) > \lambda_{\max}(A_\perp)$ and A is a $c \times c$ matrix, thus, $\lambda_{c+1}(A) = \lambda_{\max}(A_\perp)$.

By definition of EVD (eigenvalues arranged in non-increasing order) and since Λ is a $c \times c$ matrix, $\lambda_{c+1}(A + \mathcal{H}) = \lambda_{\max}(A_\perp)$. By Weyl's theorem, $\lambda_{\max}(A_\perp) = \lambda_{c+1}(A + \mathcal{H}) \leq \lambda_{c+1}(A) + \lambda_{\max}(\mathcal{H})$. Since $\lambda_{\max}(\mathcal{H}) \leq \|\mathcal{H}\|_2$, the result follows. ■

The following lemma is a minor modification of [21, Lemma 10].

Lemma 1.12: Suppose that P , \hat{P} and Q are three basis matrices, P and \hat{P} are of same size. Also, $Q'P = 0$ and $\|(I - \hat{P}\hat{P}')P\|_2 \leq \zeta_*^+$. Then,

- 1) $\|(I - \hat{P}\hat{P}')PP'\|_2 = \|(I - PP')\hat{P}\hat{P}'\|_2 = \|(I - PP')\hat{P}\|_2 = \|(I - \hat{P}\hat{P}')P\|_2 \leq \zeta_*^+$
- 2) $\|PP' - \hat{P}\hat{P}'\|_2 \leq 2\|(I - \hat{P}\hat{P}')P\|_2 \leq 2\zeta_*^+$
- 3) $\|\hat{P}'Q\|_2 \leq \zeta_*^+$
- 4) $\sqrt{1 - \zeta_*^{+2}} \leq \sigma_i((I - \hat{P}\hat{P}')Q) \leq 1$

Proof: The result follows exactly as in the proof of [21, Lemma 10]. ■

3) *Sparse Recovery Error Bound:* The following is a minor modification of [22, Theorem 1] applied to exact sparse signals.

Theorem 1.13 ([22]): Suppose we observe $y := \Psi x + z$ where z is the noise. Let \hat{x} be the solution to following problem

$$\min_x \|x\|_1 \text{ subject to } \|y - \Psi x\|_2 \leq \xi \quad (3)$$

Assume that x is s -sparse, $\|z\|_2 \leq \xi$ and $\delta_{2s}(\Psi) \leq b < (\sqrt{2} - 1)$. The solution of (3) obeys $\|\hat{x} - x\|_2 \leq C_1 \xi$ with $C_1 := \frac{4\sqrt{1+b}}{1-(\sqrt{2}+1)b}$.

II. PROBLEM DEFINITION AND MODEL ASSUMPTIONS

We give the problem definition below followed by the model and three key assumptions.

A. Problem Definition

The measurement vector at time t , M_t , is an n dimensional vector which can be decomposed as

$$M_t = L_t + S_t \quad (4)$$

Here S_t is a sparse vector with support set size at most s and minimum magnitude of nonzero values at least S_{\min} . L_t is a dense but low dimensional vector, i.e. $L_t = P_{(t)}a_t$ where $P_{(t)}$ is an $n \times r_{(t)}$ basis matrix with $r_{(t)} \ll n$, that changes every so often. $P_{(t)}$ and a_t change according to the model given below. We are given an accurate estimate of the subspace in which the initial t_{train} L_t 's lie, i.e. we are given a basis matrix \hat{P}_0 so that $\|(I - \hat{P}_0\hat{P}_0')P_0\|_2$ is small. Here P_0 is a basis matrix for $\text{span}(\mathcal{L}_{t_{\text{train}}})$, i.e. $\text{span}(P_0) = \text{span}(\mathcal{L}_{t_{\text{train}}})$. Also, for the first t_{train} time instants, S_t is either zero or very small. The goal is

- 1) to estimate both S_t and L_t at each time $t > t_{\text{train}}$, and
- 2) to estimate $\text{span}(P_{(t)})$ every-so-often, i.e., update $\hat{P}_{(t)}$ so that the subspace estimation error, $\text{SE}_{(t)} := \|(I - \hat{P}_{(t)}\hat{P}_{(t)}')P_{(t)}\|_2$ is small.

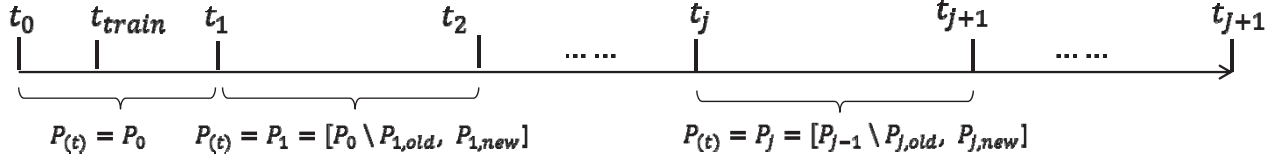


Fig. 1. The subspace change model given in Sec II-A. Here $t_0 = 0$.

Notation for S_t . Let $T_t := \{i : (S_t)_i \neq 0\}$ denote the support of S_t . Define

$$S_{\min} := \min_{t > t_{\text{train}}} \min_{i \in T_t} |(S_t)_i|, \quad \text{and} \quad s := \max_t |T_t|$$

Assumption 2.1 (Model on L_t): We assume that $L_t = P_{(t)} a_t$ where $P_{(t)}$ and a_t satisfy the following.

- 1) $P_{(t)} = P_j$ for all $t_j \leq t < t_{j+1}$, $j = 0, 1, 2, \dots, J$, where P_j is an $n \times r_j$ basis matrix with $r_j \ll n$ and $r_j \ll (t_{j+1} - t_j)$. We let $t_0 = 0$ and t_{J+1} equal the sequence length. This can be infinity also. At the change times, t_j , P_j changes as $P_j = [(P_{j-1} \setminus P_{j,\text{old}}) \ P_{j,\text{new}}]$. Here, $P_{j,\text{new}}$ is an $n \times c_{j,\text{new}}$ basis matrix with $P'_{j,\text{new}} P_{j-1} = 0$ and $P_{j,\text{old}}$ contains $c_{j,\text{old}}$ columns of P_{j-1} . Thus $r_j = r_{j-1} + c_{j,\text{new}} - c_{j,\text{old}}$. Also, $0 < t_{\text{train}} \leq t_1$. This model is illustrated in Fig. 1.
- 2) There exists a constant c_{\max} such that $0 \leq c_{j,\text{new}} \leq c_{\max}$ and $\sum_{i=1}^j (c_{i,\text{new}} - c_{i,\text{old}}) \leq c_{\max}$ for all j . Let $r_{\max} := r_0 + c_{\max}$. Thus, $r_j = r_0 + \sum_{i=1}^j (c_{i,\text{new}} - c_{i,\text{old}}) \leq r_0 + c_{\max} = r_{\max}$, i.e., the rank of P_j is upper bounded by r_{\max} .
- 3) $a_t := P_{(t)}' L_t$, is a r_j length random variable (r.v.) with the following properties.
 - a) a_t 's are mutually independent over t .
 - b) a_t is a zero mean bounded r.v., i.e. $\mathbf{E}(a_t) = 0$ and there exists a constant γ_* such that $\|a_t\|_{\infty} \leq \gamma_*$ for all t .
 - c) Its covariance matrix $\Lambda_t := \text{Cov}[a_t] = \mathbf{E}(a_t a_t')$ is diagonal with $\lambda^- := \min_t \lambda_{\min}(\Lambda_t) > 0$ and $\lambda^+ := \max_t \lambda_{\max}(\Lambda_t) < \infty$. Thus, the condition number of any Λ_t is bounded by $f := \frac{\lambda^+}{\lambda^-}$.

Also, P_j and a_t satisfy the assumptions discussed in the next three subsections.

Definition 2.2: The following notation will be used frequently. Let $P_{j,*} := P_{(t_j-1)} = P_{j-1}$. For $t \in [t_j, t_{j+1} - 1]$, let $a_{t,*} := P_{j,*}' L_t = P_{j-1}' L_t$ be the projection of L_t along $P_{j,*}$ of which $a_{t,*,\text{nz}} := (P_{j-1} \setminus P_{j,\text{old}})' L_t$ is the nonzero part. Also, let $a_{t,\text{new}} := P'_{j,\text{new}} L_t$ be the projection of L_t along the newly added directions. Thus,

$$a_{t,*} = \begin{bmatrix} a_{t,*,\text{nz}} \\ \mathbf{0} \end{bmatrix} \quad \text{and} \quad a_t = \begin{bmatrix} a_{t,*,\text{nz}} \\ a_{t,\text{new}} \end{bmatrix}$$

where $\mathbf{0}$ is a $c_{j,\text{old}}$ length zero vector (since $P_{j,\text{old}}' L_t = \mathbf{0}$). Using the above, for $t \in [t_j, t_{j+1} - 1]$, L_t can be rewritten as

$$L_t = P_j a_t = (P_{j-1} \setminus P_{j,\text{old}}) a_{t,*,\text{nz}} + P_{j,\text{new}} a_{t,\text{new}} = P_{j,*} a_{t,*} + P_{j,\text{new}} a_{t,\text{new}}$$

and Λ_t can be split as

$$\Lambda_t = \begin{bmatrix} (\Lambda_t)_{*,\text{nz}} & 0 \\ 0 & (\Lambda_t)_{\text{new}} \end{bmatrix}$$

where $(\Lambda_t)_{*,\text{nz}} := \text{Cov}(a_{t,*,\text{nz}})$ and $(\Lambda_t)_{\text{new}} = \text{Cov}(a_{t,\text{new}})$ are diagonal matrices.

B. Slow subspace change

By slow subspace change we mean all of the following.

- 1) First, the delay between consecutive subspace change times, $t_{j+1} - t_j$, is large enough.
- 2) Second, the projection of L_t along the newly added directions, $a_{t,\text{new}}$, is initially small, i.e. $\max_{t_j \leq t < t_j + \alpha} \|a_{t,\text{new}}\|_{\infty} \leq \gamma_{\text{new}}$, with $\gamma_{\text{new}} \ll \gamma_*$ and $\gamma_{\text{new}} \ll S_{\min}$, but can increase gradually. We model this as follows. Split the interval $[t_j, t_{j+1} - 1]$ into α length periods. We assume that

$$\max_j \max_{t \in [t_j + (k-1)\alpha, t_j + k\alpha - 1]} \|a_{t,\text{new}}\|_{\infty} \leq \gamma_{\text{new},k} := \min(v^{k-1} \gamma_{\text{new}}, \gamma_*)$$

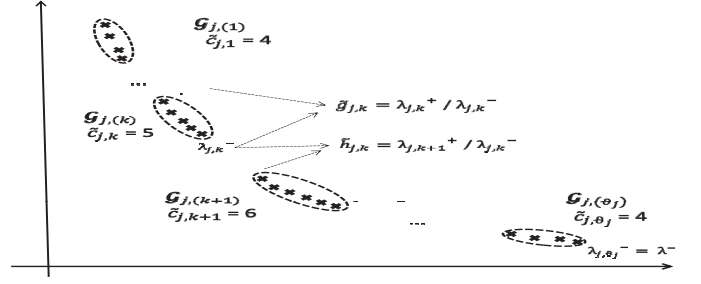


Fig. 2. We illustrate the clustering assumption. Assume $\Lambda_t = \Lambda_{\tilde{t}_j}$.

for a $v > 1$ but not too large¹. This assumption is verified for real video data in [21, Sec X-B].

3) Third, the number of newly added directions is small, i.e. $c_{j,\text{new}} \leq c_{\max} \ll r_0$. This is also verified in [21, Sec X-B].

C. Measuring denseness of a matrix and its relation with RIC

For a tall $n \times r$ matrix, B , or for a $n \times 1$ vector, B , we define the denseness coefficient as follows [21]:

$$\kappa_s(B) := \max_{|T| \leq s} \frac{\|I_T' B\|_2}{\|B\|_2}. \quad (5)$$

where $\|\cdot\|_2$ is the matrix or vector 2-norm respectively. Clearly, $\kappa_s(B) \leq 1$. As explained in [21], κ_s measures the denseness (non-compressibility) of a vector B or of the columns of a matrix B . For a vector, a small value indicates that its entries are spread out, i.e. it is a dense vector. A large value indicates that it is compressible (approximately or exactly sparse). Similarly, for an $n \times r$ matrix B , a small κ_s means that most (or all) of its columns are dense vectors.

For a basis matrix P , $\kappa_s(PP') = \kappa_s(P)$ and thus $\kappa_s(P)$ is a property of $\text{span}(P)$ [21].

Remark 2.3: A better way to quantify denseness of a matrix B would be to define the denseness coefficient as $\max_{|T| \leq s} \|I_T' Q(B)\|_2$ where $Q(B)$ is a basis matrix for $\text{span}(B)$, e.g. it can be obtained by QR decomposition on B . This definition will ensure that the denseness coefficient is a property of $\text{span}(B)$ for any matrix B . It is easy to see that $\|I_T' B\|_2 \leq \|I_T' Q(B)\|_2 \|B\|_2$. Thus, even with this new definition, all our results, and all results of [21], will go through without any change. However, we keep the definition of (5) because it was used in [21] and the current work uses certain lemmas from [21].

The following lemma was proved in [21].

Lemma 2.4: For an $n \times r$ basis matrix P (i.e P satisfying $P'P = I$),

$$\delta_s(I - PP') = \kappa_s^2(P).$$

In other words, if P is dense enough (small κ_s), then the RIC of $I - PP'$ is small. As we explain in [21, Sec IV-D], $\kappa_s(B)$ is related to the denseness assumption required by PCP [2].

D. Clustering assumption

For positive integers K and α , let $\tilde{t}_j := t_j + K\alpha$. We set their values in our main result, Theorem 4.1. Recall from the model on L_t and the slow subspace change assumption that new directions, $P_{j,\text{new}}$, get added at $t = t_j$ and initially, for the first α frames, the projection of L_t along these directions is small (and thus their variances are small), but can increase gradually. It is fair to assume that by $t = \tilde{t}_j$, the variances along these new directions have stabilized and do not change much for $t \in [\tilde{t}_j, t_{j+1} - 1]$. It is also fair to assume that the same is true for the variances along the existing directions, P_{j-1} . In other words, we assume that the matrix Λ_t is either constant or does not change much during this period. Under this assumption,

¹Small γ_{new} and slowly increasing $\gamma_{\text{new},k}$ is needed for the noise seen by the sparse recovery step to be small. However, if γ_{new} is zero or very small, it will be impossible to estimate the new subspace. This will not happen in our model because $\gamma_{\text{new}} \geq \lambda^- > 0$.

we assume that we can cluster its eigenvalues (diagonal entries) into a few clusters such that the distance between consecutive clusters is large and the distance between the smallest and largest element of each cluster is small. We make this precise below.

Assumption 2.5: Assume the following.

- 1) Either $\Lambda_t = \Lambda_{\tilde{t}_j}$ for all $t \in [\tilde{t}_j, t_{j+1} - 1]$ or Λ_t changes very little during this period so that for each $i = 1, 2, \dots, r_j$, $\min_{t \in [\tilde{t}_j, t_{j+1} - 1]} \lambda_i(\Lambda_t) \geq \max_{t \in [\tilde{t}_j, t_{j+1} - 1]} \lambda_{i+1}(\Lambda_t)$.
- 2) Let $\mathcal{G}_{j,(1)}, \mathcal{G}_{j,(2)}, \dots, \mathcal{G}_{j,(\vartheta_j)}$ be a partition of the index set $\{1, 2, \dots, r_j\}$ so that $\min_{i \in \mathcal{G}_{j,(k)}} \min_{t \in [\tilde{t}_j, t_{j+1} - 1]} \lambda_i(\Lambda_t) > \max_{i \in \mathcal{G}_{j,(k+1)}} \max_{t \in [\tilde{t}_j, t_{j+1} - 1]} \lambda_i(\Lambda_t)$, i.e. the first group/cluster contains the largest set of eigenvalues, the second one the next smallest set and so on (see Fig 2). Let
 - a) $G_{j,k} := (P_j)_{\mathcal{G}_{j,(k)}}$ be the corresponding cluster of eigenvectors, then $P_j = [G_{j,1}, G_{j,2}, \dots, G_{j,\vartheta_j}]$;
 - b) $\tilde{c}_{j,k} := |\mathcal{G}_{j,(k)}|$ be the number of elements in $\mathcal{G}_{j,(k)}$, then $\sum_{k=1}^{\vartheta_j} \tilde{c}_{j,k} = r_j$;
 - c) $\lambda_{j,k}^- := \min_{i \in \mathcal{G}_{j,(k)}} \min_{t \in [\tilde{t}_j, t_{j+1} - 1]} \lambda_i(\Lambda_t)$, $\lambda_{j,k}^+ := \max_{i \in \mathcal{G}_{j,(k)}} \max_{t \in [\tilde{t}_j, t_{j+1} - 1]} \lambda_i(\Lambda_t)$ and $\lambda_{j,\vartheta_j+1}^+ := 0$;
 - d) $\tilde{g}_{j,k} := \lambda_{j,k}^+ / \lambda_{j,k}^-$ (notice that $\tilde{g}_{j,k} \geq 1$);
 - e) $\tilde{h}_{j,k} := \lambda_{j,k+1}^+ / \lambda_{j,k}^-$ (notice that $\tilde{h}_{j,k} < 1$);
 - f) $\tilde{g}_{\max} := \max_j \max_{k=1,2,\dots,\vartheta_j} \tilde{g}_{j,k}$, $\tilde{h}_{\max} := \max_j \max_{k=1,2,\dots,\vartheta_j} \tilde{h}_{j,k}$, $\tilde{c}_{\min} := \min_j \min_{k=1,2,\dots,\vartheta_j} \tilde{c}_{j,k}$
 - g) $\vartheta_{\max} := \max_j \vartheta_j$

We assume that \tilde{g}_{\max} is small enough (the distance between the smallest and largest eigenvalues of a cluster is small) and \tilde{h}_{\max} is small enough (distance between consecutive clusters is large). We quantify this in Theorem 4.1.

Remark 2.6: The assumption above can, in fact, be relaxed to only require the following. The matrices Λ_t are such that there exists a partition, $\mathcal{G}_{j,(1)}, \mathcal{G}_{j,(2)}, \dots, \mathcal{G}_{j,(\vartheta_j)}$, of the index set $\{1, 2, \dots, r_j\}$ so that $\min_{i \in \mathcal{G}_{j,(k)}} \min_{t \in [\tilde{t}_j, t_{j+1} - 1]} \lambda_i(\Lambda_t) > \max_{i \in \mathcal{G}_{j,(k+1)}} \max_{t \in [\tilde{t}_j, t_{j+1} - 1]} \lambda_i(\Lambda_t)$. Define all quantities as above. We assume that \tilde{g}_{\max} and \tilde{h}_{\max} are small enough.

III. REPROCS WITH CLUSTER-PCA (REPROCS-CPCA)

We first briefly recap the main idea of projection-PCA (proj-PCA) which was used in [21]. The ReProCS with cluster-PCA (ReProCS-cPCA) algorithm is then explained. In Sec III-C, we discuss how to set its parameters in practice when the model may not be known. The need for proj-PCA is explained in Sec III-D. We need the following notation.

Definition 3.1: Let $\tilde{t}_j := t_j + K\alpha$. Define the following time intervals

- 1) $\mathcal{I}_{j,k} := [t_j + (k-1)\alpha, t_j + k\alpha - 1]$ for $k = 1, 2, \dots, K$.
- 2) $\tilde{\mathcal{I}}_{j,k} := [\tilde{t}_j + (k-1)\tilde{\alpha}, \tilde{t}_j + k\tilde{\alpha} - 1]$ for $k = 1, 2, \dots, \vartheta_j$.
- 3) $\tilde{\mathcal{I}}_{j,\vartheta_j+1} := [\tilde{t}_j + \vartheta_j\tilde{\alpha}, t_{j+1} - 1]$.

Notice that $[t_j, t_{j+1} - 1] = (\cup_{k=1}^K \mathcal{I}_{j,k}) \cup (\cup_{k=1}^{\vartheta_j} \tilde{\mathcal{I}}_{j,k}) \cup \tilde{\mathcal{I}}_{j,\vartheta_j+1}$. Also, K , α and $\tilde{\alpha}$ are parameters given in Algorithm 2.

A. The Projection-PCA algorithm

Given a data matrix \mathcal{D} , a basis matrix P and an integer r , projection-PCA (proj-PCA) applies PCA on $\mathcal{D}_{\text{proj}} := (I - PP')\mathcal{D}$, i.e., it computes the top r eigenvectors (the eigenvectors with the largest r eigenvalues) of $\frac{1}{\alpha_{\mathcal{D}}} \mathcal{D}_{\text{proj}} \mathcal{D}_{\text{proj}}'$. Here $\alpha_{\mathcal{D}}$ is the number of column vectors in \mathcal{D} . This is summarized in Algorithm 1.

If $P = [.]$, then projection-PCA reduces to standard PCA, i.e. it computes the top r eigenvectors of $\frac{1}{\alpha_{\mathcal{D}}} \mathcal{D} \mathcal{D}'$.

We should mention that the idea of projecting perpendicular to a partly estimated subspace has been used in different contexts in past work [36], [8].

Algorithm 1 projection-PCA: $Q \leftarrow \text{proj-PCA}(\mathcal{D}, P, r)$

- 1) Projection: compute $\mathcal{D}_{\text{proj}} \leftarrow (I - PP')\mathcal{D}$
 - 2) PCA: compute $\frac{1}{\alpha_{\mathcal{D}}} \mathcal{D}_{\text{proj}} \mathcal{D}_{\text{proj}}' \stackrel{EVD}{=} \begin{bmatrix} Q & Q_{\perp} \end{bmatrix} \begin{bmatrix} \Lambda & 0 \\ 0 & \Lambda_{\perp} \end{bmatrix} \begin{bmatrix} Q' \\ Q_{\perp}' \end{bmatrix}$ where Q is an $n \times r$ basis matrix and $\alpha_{\mathcal{D}}$ is the number of columns in \mathcal{D} .
-

Algorithm 2 Recursive Projected CS with cluster-PCA (ReProCS-cPCA)

Parameters: algorithm parameters: $\xi, \omega, \alpha, \tilde{\alpha}, K, c_{j,\text{new}}, \vartheta_j$ and $\tilde{c}_{j,i}$

Input: $n \times 1$ vector, M_t , and $n \times r_0$ basis matrix \hat{P}_0 . **Output:** $n \times 1$ vectors \hat{S}_t and \hat{L}_t , and $n \times r_{(t)}$ basis matrix $\hat{P}_{(t)}$.

Initialization: Let $\hat{P}_{(t_{\text{train}})} \leftarrow \hat{P}_0$. Let $j \leftarrow 1, k \leftarrow 1$. For $t > t_{\text{train}}$, do the following:

1) **Estimate T_t and S_t via Projected CS:**

- a) Nullify most of L_t : compute $\Phi_{(t)} \leftarrow I - \hat{P}_{(t-1)} \hat{P}'_{(t-1)}$, $y_t \leftarrow \Phi_{(t)} M_t$
- b) Sparse Recovery: compute $\hat{S}_{t,\text{cs}}$ as the solution of $\min_x \|x\|_1$ s.t. $\|y_t - \Phi_{(t)} x\|_2 \leq \xi$
- c) Support Estimate: compute $\hat{T}_t = \{i : |(\hat{S}_{t,\text{cs}})_i| > \omega\}$
- d) LS Estimate of S_t : compute $(\hat{S}_t)_{\hat{T}_t} = ((\Phi_{(t)})_{\hat{T}_t})^\dagger y_t$, $(\hat{S}_t)_{\hat{T}_t^c} = 0$

2) **Estimate L_t .** $\hat{L}_t = M_t - \hat{S}_t$.

3) **Update $\hat{P}_{(t)}$:**

- a) If $t \neq t_j + q\alpha - 1$ for any $q = 1, 2, \dots, K$ and $t \neq t_j + K\alpha + \vartheta_j \tilde{\alpha} - 1$,
 - i) set $\hat{P}_{(t)} \leftarrow \hat{P}_{(t-1)}$
 - b) **Addition: Estimate $\text{span}(P_{j,\text{new}})$ iteratively using proj-PCA:** If $t = t_j + k\alpha - 1$
 - i) $\hat{P}_{j,\text{new},k} \leftarrow \text{proj-PCA}([\hat{L}_t; t \in \mathcal{I}_{j,k}], \hat{P}_{j-1}, c_{j,\text{new}})$
 - ii) set $\hat{P}_{(t)} \leftarrow [\hat{P}_{j-1} \ \hat{P}_{j,\text{new},k}]$.
 - iii) If $k = K$, reset $k \leftarrow 1$; else increment $k \leftarrow k + 1$.
 - c) **Deletion: Estimate $\text{span}(P_j)$ by cluster-PCA:** If $t = t_j + K\alpha + \vartheta_j \tilde{\alpha} - 1$,
 - i) For $i = 1, 2, \dots, \vartheta_j$,
 - $\hat{G}_{j,i} \leftarrow \text{proj-PCA}([\hat{L}_t; t \in \tilde{\mathcal{I}}_{j,k}], [\hat{G}_{j,1}, \hat{G}_{j,2}, \dots, \hat{G}_{j,i-1}], \tilde{c}_{j,i})$
 - End for
 - ii) set $\hat{P}_j \leftarrow [\hat{G}_{j,1}, \dots, \hat{G}_{j,\vartheta_j}]$ and set $\hat{P}_{(t)} \leftarrow \hat{P}_j$.
 - iii) increment $j \leftarrow j + 1$.
-

B. The ReProCS-cPCA algorithm

ReProCS-cPCA is summarized in Algorithm 2. It proceeds as follows. The algorithm begins with the knowledge of \hat{P}_0 and initializes $\hat{P}_{(t_{\text{train}})} \leftarrow \hat{P}_0$. \hat{P}_0 can be computed as the top r_0 left singular vectors of $\mathcal{M}_{t_{\text{train}}}$ (since, by assumption, $\mathcal{S}_{t_{\text{train}}}$ is either zero or very small). For $t > t_{\text{train}}$, the following is done. Step 1 projects M_t perpendicular to $\hat{P}_{(t-1)}$, solves the ℓ_1 minimization problem, followed by support recovery and finally computes a least squares (LS) estimate of S_t on its estimated support. This final estimate \hat{S}_t is used to estimate L_t as $\hat{L}_t = M_t - \hat{S}_t$ in step 2. The sparse recovery error, $e_t := \hat{S}_t - S_t$. Since $\hat{L}_t = M_t - \hat{S}_t$, e_t also satisfies $e_t = L_t - \hat{L}_t$. Thus, a small e_t (accurate recovery of S_t) means that L_t is also recovered accurately. Step 3a is used at times when no subspace update is done. In step 3b, the estimated \hat{L}_t 's are used to obtain improved estimates of $\text{span}(P_{j,\text{new}})$ every α frames for a total of $K\alpha$ frames using the proj-PCA procedure given in Algorithm 1. As explained in [21], within K proj-PCA updates (K chosen as given in Theorem 4.1), it can be shown that both $\|e_t\|_2$ and the subspace error, $\text{SE}_{(t)} := \|(I - \hat{P}_{(t)} \hat{P}'_{(t)}) P_{(t)}\|_2$, drop down to a constant times ζ . In particular, if at $t = t_j - 1$, $\text{SE}_{(t)} \leq r\zeta$, then at $t = \tilde{t}_j := t_j + K\alpha$, we can show that $\text{SE}_{(t)} \leq (r + c_{\text{max}})\zeta$. Here $r := r_{\text{max}} = r_0 + c_{\text{max}}$.

To bring $\text{SE}_{(t)}$ down to $r\zeta$ before t_{j+1} , we need a step so that by $t = t_{j+1} - 1$ we have an estimate of only $\text{span}(P_j)$, i.e. we have “deleted” $\text{span}(P_{j,\text{old}})$. One simple way to do this is by standard PCA: at $t = \tilde{t}_j + \tilde{\alpha} - 1$, compute $\hat{P}_j \leftarrow \text{proj-PCA}([\hat{L}_t; t \in \tilde{\mathcal{I}}_{j,1}], [\cdot], r_j)$ and let $\hat{P}_{(t)} \leftarrow \hat{P}_j$. Using the $\sin \theta$ theorem and the Hoeffding corollaries, it can be shown that, as long as f is small enough, doing this is guaranteed to give an accurate estimate of $\text{span}(P_j)$. However f being small is not compatible with the slow subspace change assumption. Notice from Sec II that $\lambda^- \leq \gamma_{\text{new}}$ and $\mathbf{E}[\|L_t\|_2^2] \leq r\lambda^+$. Slow subspace change implies that γ_{new} is small. Thus, λ^- is small. However, to allow L_t to have large magnitude, λ^+ needs to be large. Thus, $f = \lambda^+/\lambda^-$ cannot be small unless we require that L_t has small magnitude for all times t .

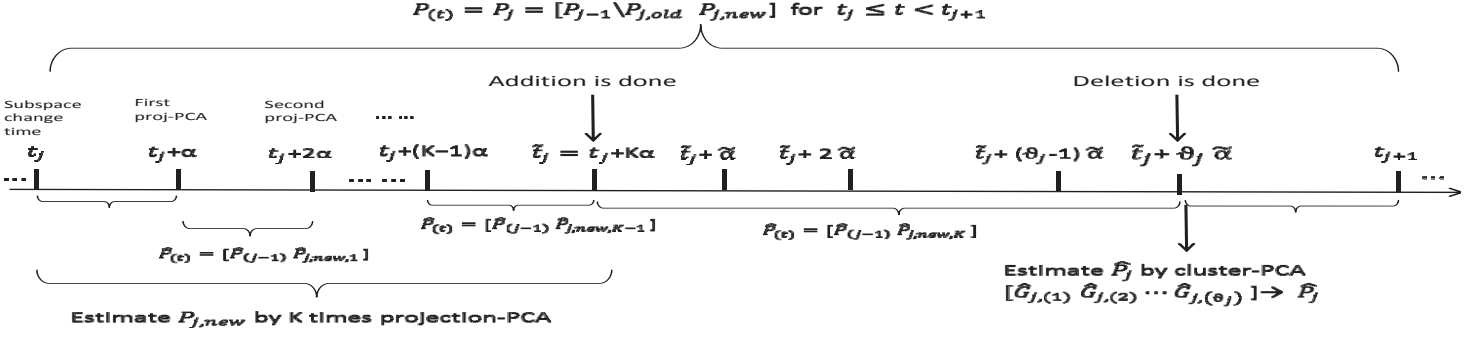


Fig. 3. A diagram illustrating subspace estimation by ReProCS-cPCA

In step 3c, we introduce a generalization of the above strategy called cluster-PCA, that removes the bound on f , but instead only requires that the eigenvalues of $\text{Cov}(L_t)$ be sufficiently clustered as explained in Sec II-D. The main idea is to recover one cluster of entries of P_j at a time. In the k^{th} iteration, we apply proj-PCA on $[\hat{L}_t; t \in \tilde{I}_{j,k}]$ with $P \leftarrow [\hat{G}_{j,1}, \hat{G}_{j,2}, \dots, \hat{G}_{j,k-1}]$ to estimate $\text{span}(G_{j,k})$. The first iteration uses $P \leftarrow []$, i.e. it computes standard PCA to estimate $\text{span}(G_{j,1})$. By modifying the approach used in [21] for analyzing the addition step, we can show that since $\tilde{g}_{j,k}$ and $\tilde{h}_{j,k}$ are small enough (by Assumption 2.5), $\text{span}(G_{j,k})$ will be accurately recovered, i.e. $\|(I - \sum_{i=1}^k \hat{G}_{j,i} \hat{G}_{j,i}') G_{j,k}\|_2 \leq \tilde{c}_{j,k} \zeta$. We do this ϑ_j times and finally we set $\hat{P}_j \leftarrow [\hat{G}_{j,1}, \hat{G}_{j,2}, \dots, \hat{G}_{j,\vartheta_j}]$ and $\hat{P}_t \leftarrow \hat{P}_j$. All of this is done at $t = \tilde{t}_j + \vartheta_j \tilde{\alpha} - 1$. Thus, at this time, $\text{SE}_t = \|(I - \hat{P}_j \hat{P}_j') P_j\|_2 \leq \sum_{k=1}^{\vartheta_j} \|(I - \sum_{i=1}^k \hat{G}_{j,i} \hat{G}_{j,i}') G_{j,k}\|_2 \leq \sum_{k=1}^{\vartheta_j} \tilde{c}_{j,k} \zeta = r_j \zeta \leq r \zeta$. Under the assumption that $t_{j+1} - t_j \geq K\alpha + \vartheta_{\max} \tilde{\alpha}$, this means that before the next subspace change time, t_{j+1} , SE_t is below $r\zeta$.

We illustrate the ideas of subspace estimation by addition proj-PCA and cluster-PCA in Fig. 3. We discuss the connection between proj-PCA done in the addition step and the cluster-PCA (for deletion) step in Table I given in Sec V-C.

C. Practical Parameter Settings

The ReProCS-cPCA algorithm has parameters $\xi, \omega, \alpha, \tilde{\alpha}, K$ and it uses knowledge of model parameters $t_j, r_0, c_{j,\text{new}}, \vartheta_j$ and $\tilde{c}_{j,i}$. If the model is known the algorithm parameters can be set as in Theorem 4.1. In practice, typically the model is unknown. In this case, the parameters $t_j, r_0, c_{j,\text{new}}, \xi, \omega, K$ can be set as explained in [21]. The parameters ϑ_j and $\tilde{c}_{j,i}$ for $i = 1, 2, \dots, \vartheta_j$, can be set by computing the eigenvalues of $\frac{1}{\alpha} \sum_{t \in \tilde{I}_{j,1}} \hat{L}_t \hat{L}_t'$ and clustering them using any standard clustering algorithm, e.g. k-means clustering or split-and-merge². We pick α and $\tilde{\alpha}$ somewhat arbitrarily. A thumb rule is that α and $\tilde{\alpha}$ need to be at least five to ten times c_{\max} and $\max_j \max_{i=1,2,\dots,\vartheta_j} \tilde{c}_{j,i}$ respectively. From simulation experiments, the algorithm is not very sensitive to the specific choice.

D. The need for Projection-PCA

The reason standard PCA cannot be used and we need proj-PCA is because $e_t = \hat{L}_t - L_t$ is correlated with L_t . The discussion here also applies to recursive or online PCA which is just a fast algorithm for computing standard PCA. In most existing works that analyze finite sample PCA, e.g. see [27] and references therein, the noise or error in the “data” used for PCA (here \hat{L}_t ’s) is uncorrelated with the true values of the data (here L_t ’s) and is zero mean. Thus, when computing the eigenvectors of $(1/\alpha) \sum_t \hat{L}_t \hat{L}_t'$, the dominant term of the perturbation, $(1/\alpha) \sum_t \hat{L}_t \hat{L}_t' - (1/\alpha) \sum_t L_t L_t'$, is $(1/\alpha) \sum_t e_t e_t'$ (the terms $(1/\alpha) \sum_t L_t e_t'$ and its transpose are close to zero w.h.p. due to law of large numbers). By assuming that the error/noise e_t is small enough, the perturbation can be made small enough.

²One simple split-and-merge approach is as follows. Start with a single cluster. Split into two clusters: select the split so that \tilde{g}_{\max} is minimized. Split each of these clusters into two parts again while ensuring \tilde{g}_{\max} is minimized. Keep doing this for d_1 steps. Notice that, with every splitting, \tilde{g}_{\max} will either remain the same or reduce, however \tilde{h}_{\max} will either remain same or increase. Then, do a set of merge steps: in each step find the pair of consecutive clusters to merge that will minimize \tilde{h}_{\max} .

However, for our problem, because e_t and L_t are correlated, the dominant terms in the perturbation seen by standard PCA will be $(1/\alpha) \sum_t L_t e_t'$ and its transpose. Since L_t can have large magnitude, the bound on the perturbation will be large and this will create problems when applying the $\sin \theta$ theorem (Theorem 1.8) to bound the subspace error. On the other hand, when using proj-PCA, L_t gets replaced by $(I - \hat{P}_{j-1} \hat{P}_{j-1}') L_t$ (in the addition step) or by $(I - \sum_{i=1}^k \hat{G}_i \hat{G}_i') L_t$ (in cluster-PCA) and this results in significantly smaller perturbation. We have explained this point in detail in Appendix F of [21].

IV. PERFORMANCE GUARANTEES

We state the main result first and then discuss it in the next subsection. We give its corollary for the case where f is small in Sec IV-C. The proof outline is given in Sec V and the proof is given in Sec VI.

A. Main Result

Theorem 4.1: Consider Algorithm 2. Let $c := c_{\max}$ and $r := r_0 + c$. Assume that L_t obeys the model given in Assumption 2.1. Also, assume that the initial subspace estimate is accurate enough, i.e. $\|(I - \hat{P}_0 \hat{P}_0') P_0\| \leq r_0 \zeta$, for a ζ that satisfies

$$\zeta \leq \min\left(\frac{10^{-4}}{(r+c)^2}, \frac{1.5 \times 10^{-4}}{(r+c)^2 f}, \frac{1}{(r+c)^3 \gamma_*^2}\right) \text{ where } f := \frac{\lambda^+}{\lambda^-}$$

Let $\xi_0(\zeta), \rho, K(\zeta), \alpha_{\text{add}}(\zeta), \alpha_{\text{del}}(\zeta), g_{j,k}$ be as defined in Definition 5.2. If the following conditions hold:

- 1) (*algorithm parameters*) $\xi = \xi_0(\zeta)$, $7\rho\xi \leq \omega \leq S_{\min} - 7\rho\xi$, $K = K(\zeta)$, $\alpha \geq \alpha_{\text{add}}(\zeta)$, $\tilde{\alpha} \geq \alpha_{\text{del}}(\zeta)$,
- 2) (*denseness*)

$$\begin{aligned} \max_j \kappa_{2s}(P_{j-1}) &\leq \kappa_{2s,*}^+ = 0.3, \quad \max_j \kappa_{2s}(P_{j,\text{new}}) \leq \kappa_{2s,\text{new}}^+ = 0.15, \\ \max_j \max_{0 \leq k \leq K} \kappa_{2s}(D_{j,\text{new},k}) &\leq \kappa_s^+ = 0.15, \quad \max_j \max_{0 \leq k \leq K} \kappa_{2s}(Q_{j,\text{new},k}) \leq \tilde{\kappa}_{2s}^+ = 0.15, \\ \max_j \kappa_s((I - \hat{P}_{j-1} \hat{P}_{j-1}' - \hat{P}_{j,\text{new},K} \hat{P}_{j,\text{new},K}') P_j) &\leq \kappa_{s,e}^+ \end{aligned}$$

where $D_{j,\text{new},k} := (I - \hat{P}_{j-1} \hat{P}_{j-1}' - \hat{P}_{j,\text{new},k} \hat{P}_{j,\text{new},k}') P_{j,\text{new}}$, and $Q_{j,\text{new},k} := (I - P_{j,\text{new}} P_{j,\text{new}}') \hat{P}_{j,\text{new},k}$ and $\hat{P}_{j,\text{new},0} = [\cdot]$,

- 3) (*slow subspace change*)

$$\begin{aligned} \max_j (t_{j+1} - t_j) &> K\alpha + \vartheta_{\max} \tilde{\alpha}, \\ \max_j \max_{t \in \mathcal{I}_{j,k}} \|a_{t,\text{new}}\|_{\infty} &\leq \gamma_{\text{new},k} := \min(1.2^{k-1} \gamma_{\text{new}}, \gamma_*), \text{ for all } k = 1, 2, \dots, K, \\ 14\rho\xi_0(\zeta) &\leq S_{\min}, \end{aligned}$$

- 4) (*small average condition number of $\text{Cov}(a_{t,\text{new}})$)* $g_{j,k} \leq g^+ := \sqrt{2}$,
- 5) (*clustered eigenvalues*) Assumption 2.5 holds with $\tilde{g}_{\max}, \tilde{h}_{\max}, \tilde{c}_{\min}$ satisfying $f_{\text{dec}}(\tilde{g}_{\max}, \tilde{h}_{\max}) - \frac{f_{\text{inc}}(\tilde{g}_{\max}, \tilde{h}_{\max})}{\tilde{c}_{\min} \zeta} > 0$ where $f_{\text{dec}}(\tilde{g}_{\max}, \tilde{h}_{\max})$ and $f_{\text{inc}}(\tilde{g}_{\max}, \tilde{h}_{\max})$ are defined in Definition 5.3 (also see Remark 7.5 which weakens this requirement),

then, with probability at least $1 - 2n^{-10}$, at all times, t ,

- 1) $\hat{T}_t = T_t$ and $\|e_t\|_2 = \|L_t - \hat{L}_t\|_2 = \|\hat{S}_t - S_t\|_2 \leq 0.18\sqrt{c}\gamma_{\text{new}} + 1.24\sqrt{\zeta}$.
- 2) the subspace error, $\text{SE}_{(t)}$ satisfies

$$\begin{aligned} \text{SE}_{(t)} &\leq \begin{cases} 0.6^{k-1} + r\zeta + 0.4c\zeta & \text{if } t \in \mathcal{I}_{j,k}, \quad k = 1, 2, \dots, K \\ (r+c)\zeta & \text{if } t \in \cup_{k=1}^{\vartheta_j} \tilde{\mathcal{I}}_{j,k} \\ r\zeta & \text{if } t \in \tilde{\mathcal{I}}_{j,\vartheta_j+1} \end{cases} \\ &\leq \begin{cases} 0.6^{k-1} + 10^{-2}\sqrt{\zeta} & \text{if } t \in \mathcal{I}_{j,k}, \quad k = 1, 2, \dots, K \\ 10^{-2}\sqrt{\zeta} & \text{if } t \in (\cup_{k=1}^{\vartheta_j} \tilde{\mathcal{I}}_{j,k}) \cup \tilde{\mathcal{I}}_{j,\vartheta_j+1} \end{cases} \end{aligned}$$

3) the error $e_t = \hat{S}_t - S_t = L_t - \hat{L}_t$ satisfies the following at various times

$$\begin{aligned} \|e_t\|_2 &\leq \begin{cases} 1.17[0.15 \cdot 0.72^{k-1} \sqrt{c}\gamma_{\text{new}} + 0.15 \cdot 0.4c\zeta \sqrt{c}\gamma_* + r\zeta \sqrt{r}\gamma_*] & \text{if } t \in \mathcal{I}_{j,k}, k = 1, 2, \dots, K \\ 1.17(r+c)\zeta \sqrt{r}\gamma_* & \text{if } t \in \bigcup_{k=1}^{\vartheta_j} \tilde{\mathcal{I}}_{j,k} \\ 1.17r\zeta \sqrt{r}\gamma_* & \text{if } t \in \tilde{\mathcal{I}}_{j,\vartheta_j+1} \end{cases} \\ &\leq \begin{cases} 0.18 \cdot 0.72^{k-1} \sqrt{c}\gamma_{\text{new}} + 1.17 \cdot 1.06\sqrt{\zeta} & \text{if } t \in \mathcal{I}_{j,k}, k = 1, 2, \dots, K \\ 1.17\sqrt{\zeta} & \text{if } t \in (\bigcup_{k=1}^{\vartheta_j} \tilde{\mathcal{I}}_{j,k}) \cup \tilde{\mathcal{I}}_{j,\vartheta_j+1} \end{cases} \end{aligned}$$

The above result says the following. Assume that the initial subspace error is small enough. If the assumptions given in the theorem hold, then, w.h.p., we will get exact support recovery at all times. Moreover, the sparse recovery error (and the error in recovering L_t) will always be bounded by $0.18\sqrt{c}\gamma_{\text{new}}$ plus a constant times $\sqrt{\zeta}$. Since ζ is very small, $\gamma_{\text{new}} \ll S_{\min}$, and c is also small, the normalized reconstruction error for S_t will be small at all times, thus making this a meaningful result. In the second conclusion, we bound the subspace estimation error, $\text{SE}_{(t)}$. When a subspace change occurs, this error is initially bounded by one. The above result shows that, w.h.p., with each addition proj-PCA step, this error decays roughly exponentially and falls below $(r+c)\zeta$ within K steps. After the cluster-PCA step, this error falls below $r\zeta$. By assumption, this occurs before the next subspace change time. Because of the choice of ζ , both $(r+c)\zeta$ and $r\zeta$ are below $0.01\sqrt{\zeta}$. The third conclusion shows that the sparse recovery error as well as the error in recovering L_t decay in a similar fashion.

B. Discussion

Notice from Definition 5.2 that $K = K(\zeta)$ is larger if ζ is smaller. Also, both $\alpha_{\text{add}}(\zeta)$ and $\alpha_{\text{del}}(\zeta)$ are inversely proportional to ζ . Thus, if we want to achieve a smaller lowest error level, ζ , we need to compute both addition proj-PCA and cluster-PCA's over larger durations, α and $\tilde{\alpha}$ respectively, and we will need more number of addition proj-PCA steps K . Because of slow subspace change, this means that we also require a larger delay between subspace change times, i.e. larger $t_{j+1} - t_j$.

1) *Comparison with ReProCS*: The ReProCS algorithm of [21] is Algorithm 2 with step 3c removed and replaced by $\hat{P}_j \leftarrow [\hat{P}_{j-1}, \hat{P}_{j,\text{new},K}]$. Let us compare the above result with that for ReProCS for the subspace change model of Assumption 2.1 [21, Corollary 43]. First, ReProCS requires $\kappa_{2s}([P_0, P_{1,\text{new}}, \dots, P_{J,\text{new}}]) \leq 0.3$ whereas ReProCS-cPCA only requires $\max_j \kappa_{2s}(P_j) \leq 0.3$. Moreover, ReProCS requires ζ to satisfy $\zeta \leq \min(\frac{10^{-4}}{(r_0+(J-1)c)^2}, \frac{1.5 \times 10^{-4}}{(r_0+(J-1)c)^2 f}, \frac{1}{(r_0+(J-1)c)^3 \gamma_*^2})$ whereas in case of ReProCS-cPCA the denominators in the bound on ζ only contain $r+c = r_0 + 2c$ (instead of $r_0 + (J-1)c$).

Because of the above, in Theorem 4.1 for ReProCS-cPCA, the only place where J (the number of subspace change times) appears is in the definitions of α_{add} and α_{del} . Notice that α_{add} and α_{del} govern the delay between subspace change times, $t_{j+1} - t_j$. Thus, with ReProCS-cPCA, J can keep increasing, as long as $t_{j+1} - t_j$ also increases accordingly. Moreover, notice that the dependence of α_{add} and α_{del} on J is only logarithmic and thus $t_{j+1} - t_j$ needs to only increase in proportion to $\log J$. On the other hand, for ReProCS (see [21, Corollary 43]), J appears in the denseness assumption, in the bound on ζ and in the definition of α_{add} . Thus, ReProCS needs a bound on J that is indirectly imposed by the denseness assumption.

The main extra assumptions that ReProCS-cPCA needs are (i) the clustering assumption (Assumption 2.5 with $\tilde{h}_{\max}, \tilde{g}_{\max}$ being small enough to satisfying $f_{\text{dec}}(\tilde{g}_{\max}, \tilde{h}_{\max}) - \frac{f_{\text{inc}}(\tilde{g}_{\max}, \tilde{h}_{\max})}{c_{\min}\zeta} > 0$; and (ii) $\max_j \kappa_s((I - \hat{P}_{j-1}\hat{P}'_{j-1} - \hat{P}_{j,\text{new},K}\hat{P}'_{j,\text{new},K})P_j) < \kappa_{s,e}^+$. The second assumption is similar to the denseness assumption on $D_{j,\text{new},k}$ which is required by both ReProCS and ReProCS-cPCA. This is discussed in [21]. The clustering assumption is a practically valid one. We verified it for a video of moving lake waters shown in <http://www.ece.iastate.edu/~chenlu/ReProCS/ReProCS.htm> as follows. We first “low-rankified” it to 90% energy as explained in [21, Sec X-B]. Note that, with one sequence, it is not possible to estimate Λ_t (this would require an ensemble of sequences) and thus it is not possible to check if all Λ_t 's in $[\tilde{t}_j, t_{j+1} - 1]$ are similar enough. However, by assuming that Λ_t is the same for a long enough sequence, one can estimate it using a time average and then verify if its eigenvalues are sufficiently clustered. When this was done, we observed that the clustering assumption holds with $\tilde{g}_{\max} = 7.2$ and $\tilde{h}_{\max} = 0.34$.

2) *Comparison with PCP*: We provide a qualitative comparison with the PCP result of [2]. A direct comparison is not possible since the proof techniques used are very different and since we solve a recursive version of the problem whereas PCP solves a batch one. Moreover, PCP provides guarantees for exact recovery of S_t and \mathcal{L}_t . In our result, we obtain guarantees for exact support recovery of the S_t 's (and hence of S_t) and bounded error recovery of its nonzero values and of \mathcal{L}_t . Also, the PCP algorithm assumes no model knowledge, whereas our algorithm does assume knowledge of model parameters. Of course, in Sec III-C, we have explained how to set the parameters in practice when the model is not known.

Consider the denseness assumptions. Let $\mathcal{L}_t = U\Sigma V'$ be its SVD. Then, for $t \in [t_j, t_{j+1} - 1]$, $U = [P_0, P_{1,\text{new}}, P_{2,\text{new}}, \dots, P_{j,\text{new}}]$ and $V = [a_1, a_2, \dots, a_t]'\Sigma^{-1}$. The result for PCP [2] assumes denseness of U and of V : it requires $\kappa_1(U) \leq \sqrt{\mu r/n}$ and $\kappa_1(V) \leq \sqrt{\mu r/n}$ for a constant $\mu \geq 1$. Moreover, it also requires $\|UV'\|_{\max} \leq \sqrt{\mu r}/n$. On the other hand, ReProCS-cPCA only requires $\kappa_{2s}(P_j) \leq 0.3$ and $\kappa_{2s}(P_{j,\text{new}}) \leq 0.15$. It does not need denseness of the entire U ; it does not assume anything about denseness of V ; and it does not need a bound on $\|UV'\|_{\max}$.

Another difference is that the result for PCP assumes that any element of the $n \times t$ matrix S_t is nonzero w.p. ϱ , and zero w.p. $1 - \varrho$, independent of all others (in particular, this means that the support sets of the different S_t 's are independent over time). Our result for ReProCS-cPCA does not put any such assumption. However it does require denseness of the matrix $D_{j,\text{new},k}$ whose columns span the unestimated part of $\text{span}(P_{j,\text{new}})$ for $t \in \mathcal{I}_{j,k+1}$. As demonstrated in Sec. VIII, this reduces ($\kappa_s(D_{j,\text{new},k})$ increases) if the support sets of S_t 's change very little over time. However, as long as, for most k , $\kappa_s(D_{j,\text{new},k})$ is anything smaller than one, which happens as long as there is at least one support change during $\mathcal{I}_{j,k}$, the subspace error does decay down to a small enough value within a finite number of steps. The number of steps required for this increases as $\kappa_s(D_{j,\text{new},k})$ increases. Since $\kappa_s(D_{j,\text{new},k})$ cannot be computed in polynomial time, for the above discussion, we computed $\|I_{T_t}' D_{j,\text{new},k}\|_2 / \|D_{j,\text{new},k}\|_2$ at $t = t_j + k\alpha - 1$ for $k = 0, 1, \dots, K$. In fact, our proof also only needs a bound on this latter quantity.

Also, some additional assumptions that ReProCS-cPCA needs are (a) accurate knowledge of the initial subspace and slow subspace change; (b) denseness of $Q_{j,\text{new},k}$; (c) the independence of a_t 's over time; (d) condition number of the average covariance matrix of $a_{t,\text{new}}$ is not too large; and (e) the clustering assumption. Assumptions (a), (b), (c) are discussed in detail in [21] and (a) is also verified for real data. As explained in [21], (c) can possibly be replaced by a weaker random walk model assumption on a_t 's if we use the matrix Azuma inequality [26] instead of matrix Hoeffding. Assumption (e) is discussed above. (d) is also an assumption made for simplicity. It can be removed if a clustering assumption similar to Assumption 2.5 holds for $(\Lambda_t)_{\text{new}} = \text{Cov}(a_{t,\text{new}})$ during $t \in [t_j, \tilde{t}_j - 1]$ and we use an approach similar to cluster-PCA. If there are $\vartheta_{\text{new},j}$ clusters, we will need $\vartheta_{\text{new},j}$ proj-PCA steps to estimate $\hat{P}_{\text{new},k}$ (instead of the current one step). At the l^{th} step, we use proj-PCA with P being \hat{P}_{j-1} concatenated with the basis matrix estimates for the last $l - 1$ clusters to recover the l^{th} cluster.

C. Special Case when f is small

If in a problem, L_t has small magnitude for all times t , then f , which is the maximum condition number of $\text{Cov}(L_t)$ for any t , can be small. If this is the case, then the clustering assumption trivially holds with $\vartheta_j = 1$, $\tilde{c}_{j,1} = r_j$, $\tilde{g}_{\max} = \tilde{g}_{j,1} = f$ and $\tilde{h}_{\max} = h_{j,1} = 0$. Thus, $\vartheta_{\max} = 1$. In this case, the following corollary holds.

Corollary 4.2: Assume that the initial subspace estimate is accurate enough as given in Theorem 4.1 with ζ as chosen there. Also assume that the first four conditions of Theorem 4.1 hold. Then, if f is small enough so that $f_{\text{inc}}(f, 0) \leq f_{\text{dec}}(f, 0)\tilde{c}_{\min}\zeta$, then, all conclusions of Theorem 4.1 hold.

Notice that the above corollary does not need Assumption 2.5 to hold.

V. DEFINITIONS, PROOF OUTLINE AND CONNECTION BETWEEN ADDITION AND DELETION STEPS

In Sec V-A, we define all the quantities that are needed for the proof. The proof outline is given in Sec V-B. We discuss how the proof strategy for the cluster-PCA (for deletion) step is related to that of addition proj-PCA in Sec V-C.

A. Definitions

Certain quantities are defined earlier in Assumptions 2.1 and 2.5, in Definitions 2.2 and 3.1, in Algorithm 2 and in Theorem 4.1.

Definition 5.1: In the sequel, we let

- 1) $c := c_{\max}$ and $r := r_{\max} = r_0 + c$ and so $r_j = r_0 + \sum_{i=1}^j (c_{i,\text{new}} - c_{i,\text{old}}) \leq r$,
- 2) $\phi^+ := 1.1735$

Definition 5.2: We define here the parameters used in Theorem 4.1.

- 1) Define $K(\zeta) := \left\lceil \frac{\log(0.6c\zeta)}{\log 0.6} \right\rceil$
- 2) Define $\xi_0(\zeta) := \sqrt{c}\gamma_{\text{new}} + 1.06\sqrt{\zeta}$
- 3) Define $\rho := \max_t \{\kappa_1(\hat{S}_{t,\text{cs}} - S_t)\}$. Notice that $\rho \leq 1$.
- 4) Define the condition number of the average of $\text{Cov}(a_{t,\text{new}})$ over $t \in \mathcal{I}_{j,k}$ as

$$g_{j,k} := \frac{\lambda_{j,\text{new},k}^+}{\lambda_{j,\text{new},k}^-} \text{ where } \lambda_{j,\text{new},k}^+ := \lambda_{\max}\left(\frac{1}{\alpha} \sum_{t \in \mathcal{I}_{j,k}} (\Lambda_t)_{\text{new}}\right), \quad \lambda_{j,\text{new},k}^- := \lambda_{\min}\left(\frac{1}{\alpha} \sum_{t \in \mathcal{I}_{j,k}} (\Lambda_t)_{\text{new}}\right),$$

- 5) Let $K = K(\zeta)$. We define $\alpha_{\text{add}}(\zeta)$ as in [21] the smallest value of α so that $(p_K(\alpha, \zeta))^{KJ} \geq 1 - n^{-10}$, where $p_K(\alpha, \zeta)$ is defined in [21, Lemma 35]. An explicit value for it [21] is

$$\alpha_{\text{add}}(\zeta) = \lceil (\log 6KJ + 11 \log n) \frac{8 \cdot 24^2}{(\zeta\lambda^-)^2} \max(\min(1.2^{4K}\gamma_{\text{new}}^4, \gamma_*^4), \frac{16}{c^2}, 4(0.186\gamma_{\text{new}}^2 + 0.0034\gamma_{\text{new}} + 2.3)^2) \rceil$$

In words, α_{add} is the smallest value of the number of data points, α , needed for an addition proj-PCA step to ensure that Theorem 4.1 holds w.p. at least $(1 - 2n^{-10})$.

- 6) We define $\alpha_{\text{del}}(\zeta)$ as the smallest value of α so that $\tilde{p}(\tilde{\alpha}, \zeta)^{\vartheta_{\max} J} \geq 1 - n^{-10}$ where $\tilde{p}(\tilde{\alpha}, \zeta)$ is defined in Lemma 7.8. We can compute an explicit value for it by using the fact that for any $x \leq 1$ and $r \geq 1$, $(1 - x)^r \geq 1 - rx$ and that $\sum_{i=1}^6 e^{-\frac{\alpha}{d_i^2}} \leq 6e^{-\frac{\alpha}{\max_{i=1,2,\dots,6} d_i^2}}$. We get

$$\alpha_{\text{del}}(\zeta) := \lceil (\log 6\vartheta_{\max} J + 11 \log n) \frac{8 \cdot 10^2}{(\zeta\lambda^-)^2} \max(4.2^2, 4b_7^2) \rceil$$

where $b_7 := (\sqrt{r}\gamma_* + \phi^+ \sqrt{\zeta})^2$ and $\phi^+ = 1.1732$. In words, α_{del} is the smallest value of the number of data points, $\tilde{\alpha}$, needed for a deletion proj-PCA step to ensure that Theorem 4.1 holds w.p. at least $(1 - 2n^{-10})$.

Definition 5.3: Define the following.

- 1) $\zeta_*^+ := r\zeta$
- 2) define the series $\{\zeta_k^+\}_{k=0,1,2,\dots,K}$ as follows

$$\zeta_0^+ := 1, \quad \zeta_k^+ := \frac{b + 0.125c\zeta}{1 - (\zeta_*^+)^2 - (\zeta_*^+)^2 f - 0.25c\zeta - b}, \text{ for } k \geq 1, \quad (6)$$

where $b := C\kappa_s^+ g^+ \zeta_{k-1}^+ + \tilde{C}(\kappa_s^+)^2 g^+ (\zeta_{k-1}^+)^2 + C' f (\zeta_*^+)^2$, $\kappa_s^+ := 0.15$, $C := (\frac{2\kappa_s^+ \phi^+}{\sqrt{1 - (\zeta_*^+)^2}} + \phi^+)$, $C' := ((\phi^+)^2 + \frac{2\phi^+}{\sqrt{1 - (\zeta_*^+)^2}} + 1 + \phi^+ + \frac{\kappa_s^+ \phi^+}{\sqrt{1 - (\zeta_*^+)^2}} + \frac{\kappa_s^+ (\phi^+)^2}{\sqrt{1 - (\zeta_*^+)^2}})$, $\tilde{C} := ((\phi^+)^2 + \frac{\kappa_s^+ (\phi^+)^2}{\sqrt{1 - (\zeta_*^+)^2}})$.

- 3) define the series $\{\tilde{\zeta}_k^+\}_{k=1,2,\dots,\vartheta_j}$ as follows

$$\tilde{\zeta}_k^+ := \frac{f_{\text{inc}}(\tilde{g}_k, \tilde{h}_k)}{f_{\text{dec}}(\tilde{g}_k, \tilde{h}_k)}$$

where $f_{\text{inc}}(\tilde{g}, \tilde{h}) := (r + c)\zeta[3\kappa_{s,e}^+ \phi^+ \tilde{g} + [\kappa_{s,e}^+ \phi^+ + \kappa_{s,e}^+ (1 + 2\phi^+) \frac{r^2 \zeta^2}{\sqrt{1 - r^2 \zeta^2}}] \tilde{h} + [\frac{r^2}{r+c} \zeta + 4r\zeta\kappa_{s,e}^+ \phi^+ + 2(r + c)\zeta(1 + \kappa_{s,e}^+ \phi^+)^2] f + 0.2\frac{1}{r+c}]$, and $f_{\text{dec}}(\tilde{g}, \tilde{h}) := 1 - \tilde{h} - 0.2\zeta - r^2 \zeta^2 f - r^2 \zeta^2 - f_{\text{inc}}(\tilde{g}, \tilde{h})$. Notice that $f_{\text{inc}}(\tilde{g}, \tilde{h})$ is an increasing function of \tilde{g}, \tilde{h} and $f_{\text{dec}}(\tilde{g}, \tilde{h})$ is a decreasing function of \tilde{g}, \tilde{h} .

As we will see, ζ_*^+ , ζ_k^+ , $\tilde{\zeta}_k^+$ are the high probability upper bounds on $\zeta_{j,*}$, $\zeta_{j,k}$, $\tilde{\zeta}_{j,k}$ (defined in Definition 5.8) under the assumptions of Theorem 4.1.

Definition 5.4: For the addition step, define

- 1) $\Phi_{j,k} := I - \hat{P}_{j-1}\hat{P}'_{j-1} - \hat{P}_{j,\text{new},k}\hat{P}'_{j,\text{new},k}$ and $\Phi_{j,0} := I - \hat{P}_{j-1}\hat{P}'_{j-1}$.
- 2) $\phi_k := \max_j \max_{T:|T|\leq s} \|((\Phi_{j,k})_T)'(\Phi_{j,k})_T)^{-1}\|_2$. It is easy to see that $\phi_k \leq \frac{1}{1 - \max_j \delta_s(\Phi_{j,k})}$.
- 3) $D_{j,\text{new},k} := \Phi_{j,k}P_{j,\text{new}}$ and $D_{j,\text{new}} := D_{j,\text{new},0} = \Phi_{j,0}P_{j,\text{new}}$.

For the cluster-PCA step (for deletion), define

- 1) $\Psi_{j,k} := I - \sum_{i=0}^k \hat{G}_{j,i}\hat{G}'_{j,i}$.
- 2) $G_{j,\text{det},k} := [G_{j,1} \cdots G_{j,k-1}]$ and $\hat{G}_{j,\text{det},k} := [\hat{G}_{j,1} \cdots \hat{G}_{j,k-1}]$. Notice that $\Psi_{j,k} = I - \hat{G}_{j,\text{det},k+1}\hat{G}'_{j,\text{det},k+1}$.
- 3) $G_{j,\text{undet},k} := [G_{j,k+1} \cdots G_{j,\vartheta_j}]$.
- 4) $D_{j,k} := \Psi_{j,k-1}G_{j,k}$, $D_{j,\text{det},k} := \Psi_{j,k-1}G_{j,\text{det},k}$ and $D_{j,\text{undet},k} := \Psi_{j,k-1}G_{j,\text{undet},k}$.

Here, $G_{j,\text{det},k}$ contains the directions that are already detected before the k^{th} step of cluster-PCA; $G_{j,k}$ contains the directions that are being detected in the current step; $G_{j,\text{undet},k}$ contains the as yet undetected directions.

Definition 5.5: Let $\kappa_{s,*} := \max_j \kappa_s(P_{j-1})$, $\kappa_{s,\text{new}} := \max_j \kappa_s(P_{j,\text{new}})$, $\kappa_{s,k} := \max_j \kappa_s(D_{j,\text{new},k})$, $\tilde{\kappa}_{s,k} := \max_j \kappa_s((I - P_{j,\text{new}}P_{j,\text{new}}')\hat{P}_{j,\text{new},k})$, $\kappa_{s,e} := \max_j \kappa_s(\Phi_K P_j)$.

Definition 5.6:

- 1) Let $D_{j,k} \stackrel{QR}{=} E_{j,k}R_{j,k}$ denote its QR decomposition. Here, $E_{j,k}$ is a basis matrix while $R_{j,k}$ is upper triangular.³
- 2) Let $E_{j,k,\perp}$ be a basis matrix for the orthogonal complement of $\text{span}(E_{j,k}) = \text{span}(D_{j,k})$. To be precise, $E_{j,k,\perp}$ is a $n \times (n - \tilde{c}_{j,k})$ basis matrix that satisfies $E_{j,k,\perp}'E_{j,k} = 0$.
- 3) Using $E_{j,k}$ and $E_{j,k,\perp}$, define $\tilde{A}_{j,k}$, $\tilde{A}_{j,k,\perp}$, $\tilde{H}_{j,k}$, $\tilde{H}_{j,k,\perp}$ and $\tilde{B}_{j,k}$ as

$$\begin{aligned} \tilde{A}_{j,k} &:= \frac{1}{\tilde{\alpha}} \sum_{t \in \tilde{I}_{j,k}} E_{j,k}' \Psi_{j,k-1} L_t L_t' \Psi_{j,k-1} E_{j,k} \\ \tilde{A}_{j,k,\perp} &:= \frac{1}{\tilde{\alpha}} \sum_{t \in \tilde{I}_{j,k}} E_{j,k,\perp}' \Psi_{j,k-1} L_t L_t' \Psi_{j,k-1} E_{j,k,\perp} \\ \tilde{H}_{j,k} &:= \frac{1}{\tilde{\alpha}} \sum_{t \in \tilde{I}_{j,k}} E_{j,k}' \Psi_{j,k-1} (e_t e_t' - L_t e_t' - e_t L_t') \Psi_{j,k-1} E_{j,k} \\ \tilde{H}_{j,k,\perp} &:= \frac{1}{\tilde{\alpha}} \sum_{t \in \tilde{I}_{j,k}} E_{j,k,\perp}' \Psi_{j,k-1} (e_t e_t' - L_t e_t' - e_t L_t') \Psi_{j,k-1} E_{j,k,\perp} \\ \tilde{B}_{j,k} &:= \frac{1}{\tilde{\alpha}} \sum_{t \in \tilde{I}_{j,k}} E_{j,k,\perp}' \Psi_{j,k-1} \hat{L}_t \hat{L}_t' \Psi_{j,k-1} E_{j,k} = \frac{1}{\tilde{\alpha}} \sum_{t \in \tilde{I}_{j,k}} E_{j,k,\perp}' \Psi_{j,k-1} (L_t - e_t)(L_t' - e_t') \Psi_{j,k-1} E_{j,k} \end{aligned}$$

4) Define

$$\begin{aligned} \tilde{\mathcal{A}}_{j,k} &:= \begin{bmatrix} E_{j,k} & E_{j,k,\perp} \end{bmatrix} \begin{bmatrix} \tilde{A}_{j,k} & 0 \\ 0 & \tilde{A}_{j,k,\perp} \end{bmatrix} \begin{bmatrix} E_{j,k}' \\ E_{j,k,\perp}' \end{bmatrix} \\ \tilde{\mathcal{H}}_{j,k} &:= \begin{bmatrix} E_{j,k} & E_{j,k,\perp} \end{bmatrix} \begin{bmatrix} \tilde{H}_{j,k} & \tilde{B}_{j,k} \\ \tilde{B}_{j,k} & \tilde{H}_{j,k,\perp} \end{bmatrix} \begin{bmatrix} E_{j,k}' \\ E_{j,k,\perp}' \end{bmatrix} \end{aligned} \quad (7)$$

5) From the above, it is easy to see that

$$\tilde{\mathcal{A}}_{j,k} + \tilde{\mathcal{H}}_{j,k} = \frac{1}{\tilde{\alpha}} \sum_{t \in \tilde{I}_{j,k}} \Psi_{j,k-1} \hat{L}_t \hat{L}_t' \Psi_{j,k-1}.$$

6) Recall from Algorithm 2 that

$$\tilde{\mathcal{A}}_{j,k} + \tilde{\mathcal{H}}_{j,k} = \frac{1}{\tilde{\alpha}} \sum_{t \in \tilde{I}_{j,k}} \Psi_{j,k-1} \hat{L}_t \hat{L}_t' \Psi_{j,k-1} \stackrel{EVD}{=} \begin{bmatrix} \hat{G}_{j,k} & \hat{G}_{j,k,\perp} \end{bmatrix} \begin{bmatrix} \Lambda_{j,k} & 0 \\ 0 & \Lambda_{j,k,\perp} \end{bmatrix} \begin{bmatrix} \hat{G}_{j,k}' \\ \hat{G}_{j,k,\perp}' \end{bmatrix}$$

³Notice that $0 < \sqrt{1 - r^2 \zeta^2} \leq \sigma_i(R_{j,k})$ by Lemma 7.3, therefore, $R_{j,k}$ is invertible.

is the EVD of $\tilde{\mathcal{A}}_{j,k} + \tilde{\mathcal{H}}_{j,k}$. Here Λ_k is a $\tilde{c}_{j,k} \times \tilde{c}_{j,k}$ diagonal matrix.

Definition 5.7: Let $\hat{P}_{j,*} := \hat{P}_{j-1} = \hat{P}_{(t_j-1)}$. Recall that $P_{j,*} := P_{(t_j-1)} = P_{j-1}$. In the sequel, we use the subscript $*$ to denote the quantity at $t = t_j - 1$.

Definition 5.8 (Subspace estimation errors):

- 1) Recall that the subspace error at time t is $SE_{(t)} := \|(I - \hat{P}_{(t)}\hat{P}_{(t)}')P_{(t)}\|_2$.
- 2) Define

$$\zeta_{j,*} := \|(I - \hat{P}_{j,*}\hat{P}_{j,*}')P_{j,*}\|_2.$$

This is the subspace error at $t = t_j - 1$, i.e. $\zeta_{j,*} = SE_{(t_j-1)}$.

- 3) For $k = 0, 1, 2, \dots, K$, define

$$\zeta_{j,k} := \|(I - \hat{P}_{j-1}\hat{P}_{j-1}' - \hat{P}_{j,\text{new},k}\hat{P}_{j,\text{new},k}')P_{j,\text{new}}\|_2.$$

This is the error in estimating $\text{span}(P_{j,\text{new}})$ after the k^{th} iteration of the addition step.

- 4) For $k = 1, 2, \dots, \vartheta_j$, define

$$\tilde{\zeta}_{j,k} := \|(I - \sum_{i=1}^k \hat{G}_{j,i}\hat{G}_{j,i}')G_{j,k}\|_2.$$

This is the error in estimating $\text{span}(G_{j,k})$ after the k^{th} iteration of the cluster-PCA step.

Remark 5.9 (Notational issue): Notice that ζ is a given scalar satisfying the bound given in Theorem 4.1, while $\zeta_{j,k}, \zeta_{j,*}$ and $\tilde{\zeta}_{j,k}$ are as defined above. Since the basis matrix estimates are functions of the \hat{L}_t 's, which in turn are depend on the L_t 's and $L_t = P_{(t)}a_t$, thus, $\zeta_{j,k}, \zeta_{j,*}$ and $\tilde{\zeta}_{j,k}$ are functions of the a_t 's. Thus, $\zeta_{j,k}, \zeta_{j,*}$ and $\tilde{\zeta}_{j,k}$ are, in fact, random variables.

Remark 5.10:

- 1) Notice that $\zeta_{j,0} = \|D_{j,\text{new}}\|_2$, $\zeta_{j,k} = \|D_{j,\text{new},k}\|_2$ and $\tilde{\zeta}_{j,k} = \|(I - \hat{G}_k\hat{G}_k')D_{j,k}\|_2 = \|\Psi_{j,k}G_{j,k}\|_2$.
- 2) Notice from the algorithm that (i) $\hat{P}_{j,\text{new},k}$ is perpendicular to $\hat{P}_{j,*} = \hat{P}_{j-1}$; and (ii) $\hat{G}_{j,k}$ is perpendicular to $[\hat{G}_{j,1}, \hat{G}_{j,2}, \dots, \hat{G}_{j,k-1}]$.
- 3) For $t \in \mathcal{I}_{j,k}$, $P_{(t)} = P_j = [(P_{j-1} \setminus P_{j,\text{old}}), P_{j,\text{new}}]$, $\hat{P}_{(t)} = [\hat{P}_{j-1}, \hat{P}_{j,\text{new},k}]$ and

$$SE_{(t)} = \|(I - \hat{P}_{j-1}\hat{P}_{j-1}' - \hat{P}_{j,\text{new},k}\hat{P}_{j,\text{new},k}')P_j\|_2 \leq \|(I - \hat{P}_{j-1}\hat{P}_{j-1}' - \hat{P}_{j,\text{new},k}\hat{P}_{j,\text{new},k}') [P_{j-1} \ P_{j,\text{new}}]\|_2 \leq \zeta_{j,*} + \zeta_{j,k}$$

for $k = 1, 2, \dots, K$. The last inequality uses the first item of this remark.

- 4) For $t \in \tilde{\mathcal{I}}_{j,k}$, $P_{(t)} = P_j$, $\hat{P}_{(t)} = [\hat{P}_{j-1}, \hat{P}_{j,\text{new},K}]$ and

$$SE_{(t)} = SE_{(t_j+K\alpha-1)} \leq \zeta_{j,*} + \zeta_{j,K}$$

- 5) For $t \in \tilde{\mathcal{I}}_{j,\vartheta_j+1}$, $P_{(t)} = P_j = [G_{j,1}, \dots, G_{j,\vartheta_j}]$, $\hat{P}_{(t)} = \hat{P}_j = [\hat{G}_{j,1}, \dots, \hat{G}_{j,\vartheta_j}]$, and

$$SE_{(t)} = \zeta_{j+1,*} \leq \sum_{k=1}^{\vartheta_j} \tilde{\zeta}_{j,k}$$

The last inequality uses the first item of this remark.

Remark 5.11: Recall that $e_t := \hat{S}_t - S_t$. Notice from Algorithm 2 that

- 1) $e_t = L_t - \hat{L}_t$.
- 2) If $\hat{T}_t = T_t$, then $e_t = I_{T_t}[(\Phi_{(t)})_{T_t}'(\Phi_{(t)})_{T_t}]^{-1}I_{T_t}'\Phi_{(t)}P_{(t)}a_t$. This follows using the definition of \hat{S}_t given in step 1d of the algorithm and the fact that $(\Phi_{(t)})_T'\Phi_{(t)} = (\Phi_{(t)}I_T)'\Phi_{(t)} = I_T'\Phi_{(t)}$ for any set T . Thus, for $t \in [t_j, t_{j+1} - 1]$,

$$e_t = I_{T_t}[(\Phi_{(t)})_{T_t}'(\Phi_{(t)})_{T_t}]^{-1}I_{T_t}'\Phi_{(t)}P_ja_t = I_{T_t}[(\Phi_{(t)})_{T_t}'(\Phi_{(t)})_{T_t}]^{-1}I_{T_t}'\Phi_{(t)}[P_{j,*}a_{t,*} + P_{j,\text{new}}a_{t,\text{new}}] \quad (8)$$

with

$$\Phi_{(t)} = \begin{cases} \Phi_{j,k-1} & t \in \mathcal{I}_{j,k}, \ k = 1, 2, \dots, K \\ \Phi_{j,K} & t \in \tilde{\mathcal{I}}_{j,k}, \ k = 1, 2, \dots, \vartheta_j \\ \Phi_{j+1,0} & t \in \tilde{\mathcal{I}}_{j,\vartheta_j+1} \end{cases}$$

TABLE I
COMPARING AND CONTRASTING THE ADDITION PROJ-PCA STEP AND PROJ-PCA USED IN THE DELETION STEP (CLUSTER-PCA)

k^{th} iteration of addition proj-PCA	k^{th} iteration of cluster-PCA in the deletion step
done at $t = t_j + k\alpha - 1$	done at $t = t_j + K\alpha + \vartheta_j\tilde{\alpha} - 1$
goal: keep improving estimates of $\text{span}(P_{j,\text{new}})$	goal: re-estimate $\text{span}(P_j)$ and thus “delete” $\text{span}(P_{j,\text{old}})$
compute $\hat{P}_{j,\text{new},k}$ by proj-PCA on $[\hat{L}_t : t \in \mathcal{I}_{j,k}]$ with $P = \hat{P}_{j-1}$	compute $\hat{G}_{j,k}$ by proj-PCA on $[\hat{L}_t : t \in \tilde{\mathcal{I}}_{j,k}]$ with $P = \hat{G}_{j,\text{det},k} = [\hat{G}_{j,1}, \dots, \hat{G}_{j,k-1}]$
start with $\ (I - \hat{P}_{j-1}\hat{P}_{j-1}')P_{j-1}\ _2 \leq r\zeta$ and $\zeta_{j,k-1} \leq \zeta_{k-1}^+ \leq 0.6^{k-1} + 0.4c\zeta$	start with $\ (I - \hat{G}_{j,\text{det},k}\hat{G}_{j,\text{det},k}')G_{j,\text{det},k}\ _2 \leq r\zeta$ and $\zeta_{j,K} \leq c\zeta$
need small $g_{j,k}$ which is the average of the condition number of $\text{Cov}(P_{j,\text{new}}'L_t)$ averaged over $t \in \mathcal{I}_{j,k}$	need small $\tilde{g}_{j,k}$ which is the maximum of the condition number of $\text{Cov}(G_{j,k}'L_t)$ over $t \in \tilde{\mathcal{I}}_{j,k}$
no undetected subspace	extra issue: ensure perturbation due to $\text{span}(G_{j,\text{undet},k})$ is small; need small $\tilde{h}_{j,k}$ to ensure the above
$\zeta_{j,k}$ is the subspace error in estimating $\text{span}(P_{j,\text{new}})$ after the k^{th} step	$\tilde{\zeta}_{j,k}$ is the subspace error in estimating $\text{span}(G_{j,k})$ after the k^{th} step
end with $\zeta_{j,k} \leq \zeta_k^+ \leq 0.6^k + 0.4c\zeta$ w.h.p.	end with $\tilde{\zeta}_{j,k} \leq \tilde{c}_{j,k}\zeta$ w.h.p.
stop when $k = K$ with K chosen so that $\zeta_{j,K} \leq c\zeta$	stop when $k = \vartheta_j$ and $\tilde{\zeta}_{j,k} \leq \tilde{c}_{j,k}\zeta$ for all $k = 1, 2, \dots, \vartheta_j$
after K^{th} iteration: $\hat{P}_{(t)} \leftarrow [\hat{P}_{j-1} \hat{P}_{j,\text{new},K}]$ and $SE_{(t)} \leq (r+c)\zeta$	after ϑ_j^{th} iteration: $\hat{P}_{(t)} \leftarrow [\hat{G}_{j,1}, \dots, \hat{G}_{j,\vartheta_j}]$ and $SE_{(t)} \leq r\zeta$

Definition 5.12: Define the random variable

$$X_{j,k_1,k_2} := \{a_1, a_2, \dots, a_{t_j+k_1\alpha+k_2\tilde{\alpha}-1}\}.$$

Recall that a_t 's are mutually independent over t .

Definition 5.13: Define the set $\tilde{\Gamma}_{j,k_1,k_2}$ as follows.

$$\tilde{\Gamma}_{j,k,0} := \{X_{j,k,0} : \zeta_{j,k} \leq \zeta_k^+, \text{ and } \hat{T}_t = T_t \text{ and } e_t \text{ satisfies (8) for all } t \in \mathcal{I}_{j,k}\}, \quad k = 1, 2, \dots, K, \quad j = 1, 2, 3, \dots, J$$

$$\tilde{\Gamma}_{j,K,k} := \{X_{j,K,k} : \tilde{\zeta}_{j,k} \leq \tilde{c}_{j,k}\zeta, \text{ and } \hat{T}_t = T_t \text{ and } e_t \text{ satisfies (8) for all } t \in \tilde{\mathcal{I}}_{j,k}\}, \quad k = 1, 2, \dots, \vartheta_j, \quad j = 1, 2, 3, \dots, J$$

$$\tilde{\Gamma}_{j,K,\vartheta_j+1} := \{X_{j+1,0,0} : \hat{T}_t = T_t \text{ and } e_t \text{ satisfies (8) for all } t \in \tilde{\mathcal{I}}_{j,\vartheta_j+1}\}, \quad j = 1, 2, 3, \dots, J$$

Define the set Γ_{j,k_1,k_2} as follows.

$$\Gamma_{1,0,0} := \{X_{1,0,0} : \zeta_{1,*} \leq r\zeta, \text{ and } \hat{T}_t = T_t \text{ and } e_t \text{ satisfies (8) for all } t \in [t_{\text{train}}, t_1 - 1]\},$$

$$\Gamma_{j,k,0} := \Gamma_{j,k-1,0} \cap \tilde{\Gamma}_{j,k,0}, \quad k = 1, 2, \dots, K, \quad j = 1, 2, 3, \dots, J$$

$$\Gamma_{j,K,k} := \Gamma_{j,K,k-1} \cap \tilde{\Gamma}_{j,K,k}, \quad k = 1, 2, \dots, \vartheta_j, \quad j = 1, 2, 3, \dots, J$$

$$\Gamma_{j+1,0,0} := \Gamma_{j,K,\vartheta_j} \cap \tilde{\Gamma}_{j,K,\vartheta_j+1}, \quad j = 1, 2, 3, \dots, J$$

Recall from the notation section that the event $\Gamma_{j,k_1,k_2}^e := \{X_{j,k_1,k_2} \in \Gamma_{j,k_1,k_2}\}$.

Remark 5.14: Notice that the subscript j always appears as the first subscript, while k is the last one. At many places in this paper, we remove the subscript j for simplicity. Whenever there is only one subscript, it refers to the value of k , e.g., Φ_0 refers to $\Phi_{j,0}$, $\hat{P}_{\text{new},k}$ refers to $\hat{P}_{j,\text{new},k}$ and so on.

B. Proof Outline of Theorem 4.1

The first part of the proof that analyzes the projected CS step and the addition step is essentially the same as that in [21]. The only difference is that, now, $\zeta_k^+ = r\zeta$ instead of $\zeta_k^+ = (r_0 + (j-1)c)\zeta$. In Lemma 6.1, the final conclusions for this part are summarized: it shows that, for all $k = 1, 2, \dots, K$, ζ_k^+ decays roughly exponentially with k and it bounds the probability of $\Gamma_{j,k,0}^e$ given $\Gamma_{j,k-1,0}^e$. The second part of the proof analyzes the projected CS step and the cluster-PCA step. The final conclusion for this part is summarized in Lemma 6.2: it bounds the probability of $\Gamma_{j,K,k}^e$ given $\Gamma_{j,K,k-1}^e$. Theorem 4.1 follows essentially by applying Lemmas 6.2 and 6.1 for each j and k and using Lemma 1.5.

Lemma 6.2, in turn, follows by combining the results of Lemma 7.2 (which shows exact support recovery and bounds the sparse recovery error for $t \in \tilde{\mathcal{I}}_{j,k}$ conditioned on $\Gamma_{j,K,k-1}^e$), and Lemma 7.8 (which bounds the subspace recovery error at

the k^{th} step of cluster-PCA conditioned on $\Gamma_{j,K,k-1}^e$). Lemma 7.2 uses the result of Lemma 7.1 which bounds the RIC of Φ_k in terms of ζ_* , ζ_k and the denseness coefficients of P_* and P_{new} . Lemma 7.8 is obtained as follows. In Lemma 7.4, we show that, under the theorem's assumptions, $\tilde{\zeta}_k^+ \leq \tilde{c}_{j,k}\zeta$. In Lemma 7.6, we bound $\tilde{\zeta}_k$ in terms of $\lambda_{\min}(A_k)$, $\lambda_{\max}(A_{k,\perp})$ and $\|\mathcal{H}_k\|_2$ using Lemma 1.11. Next, in Lemma 7.7, (i) we use Lemma 7.2 and the Hoeffding corollaries (Corollaries 1.6 and 1.7) to bound each of these terms and (ii) then we use Lemma 7.6 and these bounds to bound $\tilde{\zeta}_k$ by $\tilde{\zeta}_k^+$ with a certain probability conditioned on $\Gamma_{j,K,k-1}^e$. Finally, Lemma 7.8 follows by combining Lemma 7.4 and Lemma 7.7.

C. Connection with Addition proj-PCA

Our strategy for analyzing cluster-PCA and hence for proving Theorem 4.1 is a generalization of that used to analyze the k^{th} addition proj-PCA step in [21]. We discuss this in Table I.

VI. PROOF OF THEOREM 4.1

The theorem is a direct consequence of Lemmas 6.1 and 6.2 given below.

A. Two Main Lemmas

The lemma below is a slight modification of [21, Lemma 40]. It summarizes the final conclusions of the addition step.

Lemma 6.1 (Final lemma for addition step): Assume that all the conditions in Theorem 4.1 holds. Also assume that $\mathbf{P}(\Gamma_{j,k-1,0}^e) > 0$. Then

- 1) $\zeta_0^+ = 1$, $\zeta_k^+ \leq 0.6^k + 0.4c\zeta$ for all $k = 1, 2, \dots, K$;
- 2) $\mathbf{P}(\Gamma_{j,k,0}^e \mid \Gamma_{j,k-1,0}^e) \geq p_k(\alpha, \zeta) \geq p_K(\alpha, \zeta)$ for all $k = 1, 2, \dots, K$.

where ζ_k^+ is defined in Definition 5.3 and $p_k(\alpha, \zeta)$ is defined in [21, Lemma 35].

The proof of the above lemma follows using the exact same approach as in the proof of Lemma 40 of [21] but with $\zeta_*^+ = r\zeta$ instead of $(r_0 + (j-1)c_{\max})\zeta$ everywhere. We give the proof outline in Appendix A.

The lemma below summarizes the final conclusions for the cluster-PCA step. It is proved using lemmas given in Sec VII.

Lemma 6.2 (Final lemma for cluster-PCA): Assume that all the conditions in Theorem 4.1 hold. Also assume that $\mathbf{P}(\Gamma_{j,K,k-1}^e) > 0$. Then,

- 1) for all $k = 1, 2, \dots, \vartheta_j$, $\mathbf{P}(\Gamma_{j,K,k}^e \mid \Gamma_{j,K,k-1}^e) \geq \tilde{p}(\tilde{\alpha}, \zeta)$ where $\tilde{p}(\tilde{\alpha}, \zeta)$ is defined in Lemma 7.8;
- 2) $\mathbf{P}(\Gamma_{j+1,0,0}^e \mid \Gamma_{j,K,\vartheta_j}^e) = 1$.

Proof: Notice that $\mathbf{P}(\Gamma_{j,K,k}^e \mid \Gamma_{j,K,k-1}^e) = \mathbf{P}(\tilde{\zeta}_k \leq \tilde{c}_k\zeta \text{ and } \hat{T}_t = T_t, \text{ and } e_t \text{ satisfies (8) for all } t \in \tilde{I}_{j,k} \mid \Gamma_{j,K,k-1}^e)$ and $\mathbf{P}(\Gamma_{j+1,0,0}^e \mid \Gamma_{j,K,\vartheta_j}^e) = \mathbf{P}(\hat{T}_t = T_t \text{ and } e_t \text{ satisfies (8) for all } t \in \mathcal{I}_{j,\vartheta_j+1})$. The first claim of the lemma follows by combining Lemma 7.8 and the last claim of Lemma 7.2, both given below in Sec VII. The second claim follows using the last claim of Lemma 7.2. ■

Remark 6.3: Under the assumptions of Theorem 4.1, it is easy to see that the following holds.

- 1) For any $k = 1, 2, \dots, K$, $\Gamma_{j,k,0}^e$ implies that (i) $\zeta_{j,*} \leq \zeta_*^+ := r\zeta$ and (ii) $\zeta_{j,k'} \leq 0.6^{k'} + 0.4c\zeta$ for all $k' = 1, 2, \dots, k$
 - (i) follows from the definition of $\Gamma_{j,k,0}^e$ and $\zeta_{j,*} \leq \sum_{k=1}^{\vartheta_j-1} \tilde{\zeta}_{j-1,k'} \leq \sum_{k=1}^{\vartheta_j-1} \tilde{c}_{j-1,k'}\zeta = r_{j-1}\zeta \leq r\zeta = \zeta_*^+$; and (ii) follows from the definition of $\Gamma_{j,k,0}^e$ and the first claim of Lemma 6.1.
- 2) For any $k = 1, 2, \dots, \vartheta_j + 1$, $\Gamma_{j,K,k}^e$ implies (i) $\zeta_{j,*} \leq \zeta_*^+$, (ii) $\zeta_{j,k'} \leq 0.6^{k'} + 0.4c\zeta$ for all $k' = 1, 2, \dots, K$, (iii) $\zeta_{j,K} \leq c\zeta$, (iv) $\|\Phi_{j,K}P_j\|_2 \leq (r+c)\zeta$, (v) $\tilde{\zeta}_{j,k'} \leq \tilde{c}_{j,k'}\zeta$ for $k' = 1, 2, \dots, k$ and (vi) $\sum_{k'=1}^k \tilde{\zeta}_{j,k'} \leq r_j\zeta \leq r\zeta$.
 - (i) and (ii) follow because $\Gamma_{j,K,0}^e \subseteq \Gamma_{j,K,k}^e$, (iii) follows from (ii) using the definition of K , (iv) follows from (i) and (iii) using $\|\Phi_{j,K}P_j\|_2 \leq \|\Phi_{j,K}[P_{j,*}, P_{j,\text{new}}]\|_2 \leq \zeta_{j,*} + \zeta_{j,K}$, and (v) follows from the definition of $\Gamma_{j,K,k}^e$.
- 3) $\Gamma_{j+1,0,0}^e$ implies (i) $\zeta_{j,*} \leq \zeta_*^+$ for all j , (ii) $\zeta_{j,k} \leq 0.6^k + 0.4c\zeta$ for all $k = 1, \dots, K$ and all j , (iii) $\zeta_{j,K} \leq c\zeta$ for all j .

B. Proof of Theorem 4.1

The theorem is a direct consequence of Lemmas 6.1 and 6.2 and Lemma 1.5.

Notice that $\Gamma_{j,0,0}^e \supseteq \Gamma_{j,1,0}^e \cdots \supseteq \Gamma_{j,K,0}^e \supseteq \Gamma_{j,K,1}^e \supseteq \Gamma_{j,K,2}^e \cdots \supseteq \Gamma_{j,K,\vartheta}^e \supseteq \Gamma_{j+1,0,0}^e$. Thus, by Lemma 1.5, $\mathbf{P}(\Gamma_{j+1,0,0}^e | \Gamma_{j,0,0}^e) = \mathbf{P}(\Gamma_{j+1,0,0}^e | \Gamma_{j,K,\vartheta}^e) \prod_{k=1}^{\vartheta} \mathbf{P}(\Gamma_{j,K,k}^e | \Gamma_{j,K,k-1}^e) \prod_{k=1}^K \mathbf{P}(\Gamma_{j,k,0}^e | \Gamma_{j,k-1,0}^e)$ and $\mathbf{P}(\Gamma_{j+1,0,0}^e | \Gamma_{1,0,0}^e) = \prod_{j=1}^J \mathbf{P}(\Gamma_{j+1,0,0}^e | \Gamma_{j,0,0}^e)$. Using Lemmas 6.1 and 6.2, and the fact that $p_k(\alpha, \zeta) \geq p_K(\alpha, \zeta)$ [21, Lemma 35], we get $\mathbf{P}(\Gamma_{j+1,0,0}^e | \Gamma_{1,0,0}^e) \geq p_K(\alpha, \zeta)^{KJ} \tilde{p}(\tilde{\alpha}, \zeta)^{\vartheta_{\max} J}$. Also, $\mathbf{P}(\Gamma_{1,0,0}^e) = 1$. This follows by the assumption on \hat{P}_0 and Lemma 7.2. Thus, $\mathbf{P}(\Gamma_{j+1,0,0}^e) \geq p_K(\alpha, \zeta)^{KJ} \tilde{p}(\tilde{\alpha}, \zeta)^{\vartheta_{\max} J}$.

Using the definitions of $\alpha_{\text{add}}(\zeta)$ and $\alpha_{\text{del}}(\zeta)$ and $\alpha \geq \alpha_{\text{add}}$ and $\tilde{\alpha} \geq \alpha_{\text{del}}$, $\mathbf{P}(\Gamma_{j+1,0,0}^e) \geq p_K(\alpha, \zeta)^{KJ} \tilde{p}(\tilde{\alpha}, \zeta)^{\vartheta_{\max} J} \geq (1 - n^{-10})^2 \geq 1 - 2n^{-10}$.

The event $\Gamma_{j+1,0,0}^e$ implies that $\hat{T}_t = T_t$ and e_t satisfies (8) for all $t < t_{J+1}$. Using Remark 5.10 and the third claim of Remark 6.3, $\Gamma_{j+1,0,0}^e$ implies that all the bounds on the subspace error hold. Using these, Remark 5.11, $\|a_{t,\text{new}}\|_2 \leq \sqrt{c}\gamma_{\text{new},k}$ and $\|a_t\|_2 \leq \sqrt{r}\gamma_*$, $\Gamma_{j+1,0,0}^e$ implies that all the bounds on $\|e_t\|_2$ hold (the bounds are obtained in in Lemmas 7.2 and A.2).

Thus, all conclusions of the the result hold w.p. at least $1 - 2n^{-10}$.

VII. LEMMAS USED TO PROVE LEMMA 6.2

In this section, we remove the subscript j at most places. The convention of Remark 5.14 applies.

A. Showing exact support recovery and getting an expression for e_t

Lemma 7.1 (Bounding the RIC of Φ_k): The following hold.

- 1) $\delta_s(\Phi_0) = \kappa_s^2(\hat{P}_*) \leq \kappa_{s,*}^2 + 2\zeta_*$
- 2) $\delta_s(\Phi_k) = \kappa_s^2([\hat{P}_* \hat{P}_{\text{new},k}]) \leq \kappa_s^2(\hat{P}_*) + \kappa_s^2(\hat{P}_{\text{new},k}) \leq \kappa_{s,*}^2 + 2\zeta_* + (\kappa_{s,\text{new}} + \tilde{\kappa}_{s,k}\zeta_k + \zeta_*)^2$ for $k = 1, 2 \dots K$

Proof: The above lemma is the same as the last two claims of [21, Lemma 28]. It follows using Lemma 2.4 and some linear algebraic manipulations. ■

Lemma 7.2 (Sparse recovery, support recovery and expression for e_t): Assume that the conditions of Theorem 4.1 hold.

- 1) For all $k = 1, 2, \dots, \vartheta + 1$, $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$ implies that
 - a) $\zeta_* \leq \zeta_*^+ := r\zeta$, $\zeta_K \leq c\zeta$, $\|\Phi_K P_j\|_2 \leq (r+c)\zeta$,
 - b) $\delta_s(\Phi_K) \leq 0.1479$ and $\phi_K \leq \phi^+ := 1.1735$
 - c) for any $t \in \tilde{\mathcal{I}}_{j,k}$,
 - i) the projection noise $\beta_t := (I - \hat{P}_{(t-1)}\hat{P}'_{(t-1)})L_t$ satisfies $\|\beta_t\|_2 \leq \sqrt{\zeta}$,
 - ii) the CS error satisfies $\|\hat{S}_{t,\text{cs}} - S_t\|_2 \leq 7\sqrt{\zeta}$,
 - iii) $\hat{T}_t = T_t$,
 - iv) e_t satisfies (8) and $\|e_t\|_2 \leq \phi^+ \sqrt{\zeta}$.
- 2) For all $k = 1, 2, \dots, \vartheta + 1$, $\mathbf{P}(T_t = \hat{T}_t \text{ and } e_t \text{ satisfies (8) for all } t \in \tilde{\mathcal{I}}_{j,k} | X_{j,K,k-1}) = 1$ for all $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$.
- 3) For all $k = 1, 2, \dots, \vartheta + 1$, $\mathbf{P}(T_t = \hat{T}_t \text{ and } e_t \text{ satisfies (8) for all } t \in \tilde{\mathcal{I}}_{j,k} | \Gamma_{j,K,k-1}^e) = 1$.

Proof:

Claim 1-a follows using Remark 6.3. Claim 1-b) follows using claim 1-a) and Lemma 7.1. Claim 1-c) follows in a fashion similar to the proof of [21, Lemma 30]. The main difference is that everywhere we use $\Phi_K L_t = \Phi_K P_j a_t$ and $\|\Phi_K P_j\|_2 \leq (r+c)\zeta$. Claim 1-c-i) uses this and the fact that for $t \in \tilde{\mathcal{I}}_{j,k}$, $\Phi_{(t)} = \Phi_K$, and $\sqrt{\zeta} \leq \sqrt{\gamma_*^2/(r+c)^3}$. Claim 1-c-ii) uses c-i), $\sqrt{\zeta} \leq \xi$ (defined in the theorem), $\delta_{2s}(\Phi_K) \leq 0.1479$, and Theorem 1.13. Claim 1-c-iii) uses c-ii), the definition of ρ , the choice of ω and the lower bound on S_{\min} given in the theorem. Claim 1-c-iv) uses claim c-iii) and Remark 5.11. To get the bound on $\|e_t\|_2$ we use the first expression of (8), $\phi_K \leq \phi^+ := 1.1735$, and $\sqrt{\zeta} \leq \sqrt{\gamma_*^2/(r+c)^3}$.

Claim 2) is just a rewrite of claim 1). Claim 3) follows from claim 2) by Lemma 1.4. ■

B. A lemma needed for bounding the subspace error, $\tilde{\zeta}_k$

Lemma 7.3: Assume that $\tilde{\zeta}_{k'} \leq \tilde{c}_{k'} \zeta$ for $k' = 1, \dots, k-1$. Then

- 1) $\|D_{\det,k}\|_2 = \|\Psi_{k-1} G_{\det,k}\|_2 \leq r\zeta$.
- 2) $\|G_{\det,k} G_{\det,k}' - \hat{G}_{\det,k} \hat{G}_{\det,k}'\|_2 \leq 2r\zeta$.
- 3) $0 < \sqrt{1-r^2\zeta^2} \leq \sigma_i(D_k) = \sigma_i(R_k) \leq 1$. Thus, $\|D_k\|_2 = \|R_k\|_2 \leq 1$ and $\|D_k^{-1}\|_2 = \|R_k^{-1}\|_2 \leq 1/\sqrt{1-r^2\zeta^2}$.
- 4) $\|D_{\det,k}' E_k\|_2 = \|G_{\det,k}' E_k\|_2 \leq \frac{r^2\zeta^2}{\sqrt{1-r^2\zeta^2}}$.

Proof: The first claim essentially follows by using the fact that $\hat{G}_1, \dots, \hat{G}_{k-1}$ are mutually orthonormal and triangle inequality. Recall that $\Psi_{k-1} = (I - \hat{G}_{\det,k} \hat{G}_{\det,k}')$. The last three claims use this and the first claim and apply Lemma 1.12. The last claim also uses the definition of D_k and its QR decomposition. The complete proof is given in Appendix B. ■

C. Bounding on the subspace error, $\tilde{\zeta}_k$

Lemma 7.4 (Bounding $\tilde{\zeta}_k^+$): If

$$f_{dec}(\tilde{g}_{\max}, \tilde{h}_{\max}) - \frac{f_{inc}(\tilde{g}_{\max}, \tilde{h}_{\max})}{\tilde{c}_{\min}\zeta} > 0 \quad (9)$$

then $f_{dec}(\tilde{g}_k, \tilde{h}_k) > 0$ and $\tilde{\zeta}_k^+ \leq \tilde{c}_k \zeta$.

Proof: Recall that $f_{inc}(\cdot)$, $f_{dec}(\cdot)$ are defined in Definition 5.3 and $\tilde{\zeta}_k^+ := \frac{f_{inc}(\tilde{g}, \tilde{h})}{f_{dec}(\tilde{g}, \tilde{h})}$. Notice that $f_{inc}(\cdot)$ is a non-decreasing function of \tilde{g}, \tilde{h} , and $f_{dec}(\cdot)$ is a non-increasing function. Using the definition of $\tilde{g}_{\max}, \tilde{h}_{\max}, \tilde{c}_{\min}$ given in Assumption 2.5, the result follows. ■

Remark 7.5: If we ignore the small terms of $f_{inc}(\cdot)$ and $f_{dec}(\cdot)$, the above condition simplifies to requiring that $\frac{3\kappa_{s,e}^+ \phi^+ \tilde{g}_{\max} + \kappa_{s,e}^+ \phi^+ \tilde{h}_{\max}}{1 - \tilde{h}_{\max}} \leq \frac{\tilde{c}_{\min}}{r + c}$. Since $\tilde{g}_{\max} \geq 1$, the first term of the numerator is the largest one. To ensure that this condition holds we need $\kappa_{s,e}^+$ to be very small. However, as explained in Sec VII-D, if we also assume denseness of D_k , i.e. if we assume $\kappa_s(D_k) \leq \kappa_{s,D}^+$ for a small enough $\kappa_{s,D}^+$, then the first term of the numerator can be replaced by $\max(3\kappa_{s,e}^+ \kappa_{s,D}^+ \phi^+ \tilde{g}_{\max}, \kappa_{s,e}^+ \phi^+ \tilde{h}_{\max})$. This will relax the requirement on $\kappa_{s,e}^+$, e.g. now $\kappa_{s,e}^+ = \kappa_{s,D}^+ = 0.3$ will work.

Lemma 7.6 (Bounding $\tilde{\zeta}_k$): If $\lambda_{\min}(\tilde{A}_k) - \lambda_{\max}(\tilde{A}_{k,\perp}) - \|\tilde{\mathcal{H}}_k\|_2 > 0$, then

$$\tilde{\zeta}_k \leq \frac{\|\tilde{\mathcal{H}}_k\|_2}{\lambda_{\min}(\tilde{A}_k) - \lambda_{\max}(\tilde{A}_{k,\perp}) - \|\tilde{\mathcal{H}}_k\|_2} \quad (10)$$

Proof: Recall that $\tilde{A}_k, \tilde{A}_{k,\perp}, \tilde{\mathcal{H}}_k$ are defined in Definition 5.6. The result follows by using the fact that $\tilde{\zeta}_k = \|(I - \hat{G}_k \hat{G}_k') D_{j,k}\|_2 = \|(I - \hat{G}_k \hat{G}_k') E_k R_k\|_2 \leq \|(I - \hat{G}_k \hat{G}_k') E_k\|_2$ and applying Lemma 1.11 with $E \equiv E_k$ and $F \equiv \hat{G}_k$. ■

Lemma 7.7 (High probability bounds for each of the terms in the $\tilde{\zeta}_k$ bound and for $\tilde{\zeta}_k$): Assume that the conditions of Theorem 4.1 hold. Also, assume that $\mathbf{P}(\Gamma_{j,K,k-1}^e) > 0$. Then, for all $1 \leq k \leq \vartheta_j$,

- 1) $\mathbf{P}(\lambda_{\min}(\tilde{A}_k) \geq \lambda_k^-(1 - r^2\zeta^2 - 0.1\zeta) | \Gamma_{j,K,k-1}^e) > 1 - \tilde{p}_1(\tilde{\alpha}, \zeta)$ with $\tilde{p}_1(\tilde{\alpha}, \zeta)$ given in (14).
- 2) $\mathbf{P}(\lambda_{\max}(\tilde{A}_{k,\perp}) \leq \lambda_k^-(\tilde{h}_k + r^2\zeta^2 f + 0.1\zeta) | \Gamma_{j,K,k-1}^e) > 1 - \tilde{p}_2(\tilde{\alpha}, \zeta)$ with $\tilde{p}_2(\tilde{\alpha}, \zeta)$ given in (15).
- 3) $\mathbf{P}(\|\tilde{\mathcal{H}}_k\|_2 \leq \lambda_k^- f_{inc}(\tilde{g}_k, \tilde{h}_k) | \Gamma_{j,K,k-1}^e) \geq 1 - \tilde{p}_3(\tilde{\alpha}, \zeta)$ with $\tilde{p}_3(\tilde{\alpha}, \zeta)$ given in (20).
- 4) $\mathbf{P}(\lambda_{\min}(\tilde{A}_k) - \lambda_{\max}(\tilde{A}_{k,\perp}) - \|\tilde{\mathcal{H}}_k\|_2 \geq \lambda_k^- f_{dec}(\tilde{g}_k, \tilde{h}_k) | \Gamma_{j,K,k-1}^e) \geq \tilde{p}(\tilde{\alpha}, \zeta) := 1 - \tilde{p}_1(\tilde{\alpha}, \zeta) - \tilde{p}_2(\tilde{\alpha}, \zeta) - \tilde{p}_3(\tilde{\alpha}, \zeta)$.
- 5) If $f_{dec}(\tilde{g}_k, \tilde{h}_k) > 0$, then $\mathbf{P}(\tilde{\zeta}_k \leq \tilde{\zeta}_k^+ | \Gamma_{j,K,k-1}^e) \geq \tilde{p}(\tilde{\alpha}, \zeta)$

Proof: Recall that $f_{inc}(\cdot)$, $f_{dec}(\cdot)$ and $\tilde{\zeta}_k^+$ are defined in Definition 5.3. The proof of the first three claims is given in Sec VII-D. The fourth claim follows directly from the first three using the union bound on probabilities. The fifth claim follows from the fourth using Lemma 7.6. ■

Lemma 7.8 (High probability bound on $\tilde{\zeta}_k$): Assume that the conditions of Theorem 4.1 hold. Then,

$$\mathbf{P}(\tilde{\zeta}_k \leq \tilde{c}_k \zeta | \Gamma_{j,K,k-1}^e) \geq \tilde{p}(\tilde{\alpha}, \zeta)$$

Proof: This follows by combining Lemma 7.4 and the last claim of Lemma 7.7. ■

D. Proof of Lemma 7.7

Proof: We use $\frac{1}{\alpha} \sum_t$ to denote $\frac{1}{\alpha} \sum_{t \in \tilde{\mathcal{I}}_{j,k}}$.

For $t \in \tilde{\mathcal{I}}_{j,k}$, let $a_{t,k} := G_{j,k}' L_t$, $a_{t,\det} := G_{\det,k}' L_t = [G_{j,1}, \dots, G_{j,k-1}]' L_t$ and $a_{t,\text{undet}} := G_{\text{undet},k}' L_t = [G_{j,k+1}, \dots, G_{j,\vartheta_j}]' L_t$. Then $a_t := P_j' L_t$ can be split as $a_t = [a_{t,\det}' \ a_{t,k}' \ a_{t,\text{undet}}']'$.

This lemma follows using the following facts and the Hoeffding corollaries, Corollary 1.6 and 1.7.

- 1) The statement “conditioned on r.v. X , the event \mathcal{E}^e holds w.p. one for all $X \in \Gamma$ ” is equivalent to “ $\mathbf{P}(\mathcal{E}^e | X) = 1$, for all $X \in \Gamma$ ”. We often use the former statement in our proofs since it is often easier to interpret.
- 2) The matrices D_k , R_k , E_k , $D_{\det,k}$, $D_{\text{undet},k}$, Ψ_{k-1} , Φ_K are functions of the r.v. $X_{j,K,k-1}$. All terms that we bound for the first two claims of the lemma are of the form $\frac{1}{\alpha} \sum_{t \in \tilde{\mathcal{I}}_{j,k}} Z_t$ where $Z_t = f_1(X_{j,K,k-1}) Y_t f_2(X_{j,K,k-1})$, Y_t is a sub-matrix of $a_t a_t'$ and $f_1(\cdot)$ and $f_2(\cdot)$ are functions of $X_{j,K,k-1}$. For instance, one of the terms while bounding $\lambda_{\min}(\mathcal{A}_k)$ is $\frac{1}{\alpha} \sum_t R_k a_{t,k} a_{t,k}' R_k'$.
- 3) $X_{j,K,k-1}$ is independent of any a_t for $t \in \tilde{\mathcal{I}}_{j,k}$, and hence the same is true for the matrices D_k , R_k , E_k , $D_{\det,k}$, $D_{\text{undet},k}$, Ψ_{k-1} , Φ_K . Also, a_t 's for different $t \in \tilde{\mathcal{I}}_{j,k}$ are mutually independent. Thus, conditioned on $X_{j,K,k-1}$, the Z_t 's defined above are mutually independent.
- 4) All the terms that we bound for the third claim contain e_t . Using the second claim of Lemma 7.2, conditioned on $X_{j,K,k-1}$, e_t satisfies (8) w.p. one whenever $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$. Conditioned on $X_{j,K,k-1}$, all these terms are also of the form $\frac{1}{\alpha} \sum_{t \in \tilde{\mathcal{I}}_{j,k}} Z_t$ with Z_t as defined above, whenever $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$. Thus, conditioned on $X_{j,K,k-1}$, the Z_t 's for these terms are mutually independent, whenever $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$.
- 5) By Remark 6.3, $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$ implies that $\zeta_* \leq r\zeta$, $\tilde{\zeta}_{k'} \leq c_{k'}\zeta$, for all $k' = 1, 2, \dots, k-1$, $\zeta_K \leq \zeta_K^+ \leq c\zeta$, (iv) $\phi_K \leq \phi^+$ (by Lemma 7.2); (v) $\|\Phi_K P_j\|_2 \leq (r+c)\zeta$; and (vi) all conclusions of Lemma 7.3 hold.
- 6) By the clustering assumption, $\lambda_k^- \leq \lambda_{\min}(\mathbf{E}(a_{t,k} a_{t,k}')) \leq \lambda_{\max}(\mathbf{E}(a_{t,k} a_{t,k}')) \leq \lambda_k^+$; $\lambda_{\max}(\mathbf{E}(a_{t,\det} a_{t,\det}')) \leq \lambda_1^+ = \lambda^+$; and $\lambda_{\max}(\mathbf{E}(a_{t,\text{undet}} a_{t,\text{undet}}')) \leq \lambda_{k+1}^+$. Also, $\lambda_{\max}(\mathbf{E}(a_t a_t')) \leq \lambda^+$.
- 7) By Weyl's theorem, for a sequence of matrices B_t , $\lambda_{\min}(\sum_t B_t) \geq \sum_t \lambda_{\min}(B_t)$ and $\lambda_{\max}(\sum_t B_t) \leq \sum_t \lambda_{\max}(B_t)$.

Consider $\tilde{A}_k = \frac{1}{\alpha} \sum_t E_k' \Psi_{k-1} L_t L_t' \Psi_{k-1} E_k$. Notice that $E_k' \Psi_{k-1} L_t = R_k a_{t,k} + E_k' (D_{\det,k} a_{t,\det} + D_{\text{undet},k} a_{t,\text{undet}})$. Let $Z_t = R_k a_{t,k} a_{t,k}' R_k'$ and let $Y_t = R_k a_{t,k} (a_{t,\det}' D_{\det,k}' + a_{t,\text{undet}}' D_{\text{undet},k}') E_k + E_k' (D_{\det,k} a_{t,\det} + D_{\text{undet},k} a_{t,\text{undet}}) a_{t,k}' R_k'$. Then

$$\tilde{A}_k \succeq \frac{1}{\alpha} \sum_t Z_t + \frac{1}{\alpha} \sum_t Y_t \quad (11)$$

Consider $\frac{1}{\alpha} \sum_t Z_t = \frac{1}{\alpha} \sum_t R_k a_{t,k} a_{t,k}' R_k'$. (a) As explained above, the Z_t 's are conditionally independent given $X_{j,K,k-1}$. (b) Using Ostrowski's theorem and Lemma 7.3, for all $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$, $\lambda_{\min}(\mathbf{E}(\frac{1}{\alpha} \sum_t Z_t | X_{j,K,k-1})) = \lambda_{\min}(R_k \frac{1}{\alpha} \sum_t \mathbf{E}(a_{t,k} a_{t,k}') R_k') \geq \lambda_{\min}(R_k R_k') \lambda_{\min}(\frac{1}{\alpha} \sum_t \mathbf{E}(a_{t,k} a_{t,k}')) \geq (1 - r^2 \zeta^2) \lambda_k^-$. (c) Finally, using $\|R_k\|_2 \leq 1$ and $\|a_{t,k}\|_2 \leq \sqrt{\tilde{c}_k} \gamma_*$, conditioned on $X_{j,K,k-1}$, $0 \preceq Z_t \preceq \tilde{c}_k \gamma_*^2 I$ holds w.p. one for all $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$.

Thus, applying Corollary 1.6 with $\epsilon = 0.1\zeta\lambda^-$, and using $\tilde{c}_k \leq r$, for all $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$,

$$\mathbf{P}(\lambda_{\min}(\frac{1}{\alpha} \sum_t Z_t) \geq (1 - r^2 \zeta^2) \lambda_k^- - 0.1\zeta\lambda^- | X_{j,K,k-1}) \geq 1 - \tilde{c}_k \exp(-\frac{\tilde{\alpha}\epsilon^2}{8(\tilde{c}_k \gamma_*^2)^2}) \geq 1 - r \exp(-\frac{\tilde{\alpha} \cdot (0.1\zeta\lambda^-)^2}{8r^2 \gamma_*^4}) \quad (12)$$

Consider $Y_t = R_k a_{t,k} (a_{t,\det}' D_{\det,k}' + a_{t,\text{undet}}' D_{\text{undet},k}') E_k + E_k' (D_{\det,k} a_{t,\det} + D_{\text{undet},k} a_{t,\text{undet}}) a_{t,k}' R_k'$. (a) As before, the Y_t 's are conditionally independent given $X_{j,K,k-1}$. (b) Since $\mathbf{E}[a_t] = 0$ and $\text{Cov}[a_t] = \Lambda_t$ is diagonal, $\mathbf{E}(\frac{1}{\alpha} \sum_t Y_t | X_{j,K,k-1}) = 0$ whenever $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$. (c) Conditioned on $X_{j,K,k-1}$, $\|Y_t\|_2 \leq 2\sqrt{\tilde{c}_k} r \gamma_*^2 \zeta (1 + \frac{r\zeta}{\sqrt{1-r^2\zeta^2}}) \leq 2r^2 \zeta \gamma_*^2 (1 + \frac{10^{-4}}{\sqrt{1-10^{-4}}}) \leq \frac{2}{r} (1 + \frac{10^{-4}}{\sqrt{1-10^{-4}}}) < 2.1$ holds w.p. one for all $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$. This follows because $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$ implies that $\|D_{\det,k}\|_2 \leq r\zeta$, $\|E_k' D_{\text{undet},k}\|_2 = \|E_k' G_{\text{undet},k}\|_2 \leq \frac{r^2 \zeta^2}{\sqrt{1-r^2 \zeta^2}}$. Thus, under the same conditioning, $-bI \preceq Y_t \preceq bI$ with $b = 2.1$ w.p. one. Thus, applying Corollary 1.6 with $\epsilon = 0.1\zeta\lambda^-$, we get

$$\mathbf{P}(\lambda_{\min}(\frac{1}{\alpha} \sum_t Y_t) \geq -0.1\zeta\lambda^- | X_{j,K,k-1}) \geq 1 - r \exp(-\frac{\tilde{\alpha}(0.1\zeta\lambda^-)^2}{8(4.2)^2}) \text{ for all } X_{j,K,k-1} \in \Gamma_{j,K,k-1} \quad (13)$$

Combining (11), (12) and (13) and using the union bound, $\mathbf{P}(\lambda_{\min}(\tilde{A}_k) \geq \lambda_k^-(1 - r^2\zeta^2) - 0.2\zeta\lambda^- | X_{j,K,k-1}) \geq 1 - \tilde{p}_1(\tilde{\alpha}, \zeta)$ for all $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$ where

$$\tilde{p}_1(\tilde{\alpha}, \zeta) := r \exp\left(-\frac{\tilde{\alpha} \cdot (0.1\zeta\lambda^-)^2}{8r^2\gamma_*^4}\right) + r \exp\left(-\frac{\tilde{\alpha}(0.1\zeta\lambda^-)^2}{8(4.2)^2}\right) \quad (14)$$

The first claim of the lemma follows by using $\lambda_k^- \geq \lambda^-$ and applying Lemma 1.4 with $X \equiv X_{j,K,k-1}$ and $\mathcal{C} \equiv \Gamma_{j,K,k-1}$.

Consider $\tilde{A}_{k,\perp} := \frac{1}{\alpha} \sum_t E_{k,\perp}' \Psi_{k-1} L_t L_t' \Psi_{k-1} E_{k,\perp}$. Notice that $E_{k,\perp}' \Psi_{k-1} L_t = E_{k,\perp}' (D_{\det,k} a_{t,\det} + D_{\text{undet},k} a_{t,\text{undet}})$. Thus, $\tilde{A}_{k,\perp} = \frac{1}{\alpha} \sum_t Z_t$ with $Z_t = E_{k,\perp}' (D_{\det,k} a_{t,\det} + D_{\text{undet},k} a_{t,\text{undet}}) (D_{\det,k} a_{t,\det} + D_{\text{undet},k} a_{t,\text{undet}})' E_{k,\perp}$ which is of size $(n - \tilde{c}_k) \times (n - \tilde{c}_k)$. (a) As before, given $X_{j,K,k-1}$, the Z_t 's are independent. (b) Conditioned on $X_{j,K,k-1}$, $0 \preceq Z_t \preceq r\gamma_*^2 I$ w.p. one for all $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$. (c) $\mathbf{E}(\frac{1}{\alpha} \sum_t Z_t | X_{j,K,k-1}) \preceq (\lambda_{k+1}^+ + r^2\zeta^2\lambda^+) I$ for all $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$.

Thus applying Corollary 1.6 with $\epsilon = 0.1\zeta\lambda^-$ and using $\tilde{c}_k \geq \tilde{c}_{\min}$, we get

$$\mathbf{P}(\lambda_{\max}(\tilde{A}_{k,\perp}) \leq \lambda_{k+1}^+ + r^2\zeta^2\lambda^+ + 0.1\zeta\lambda^- | X_{j,K,k-1}) \geq 1 - \tilde{p}_2(\tilde{\alpha}, \zeta) \text{ for all } X_{j,K,k-1} \in \Gamma_{j,K,k-1}$$

where

$$\tilde{p}_2(\tilde{\alpha}, \zeta) := (n - \tilde{c}_{\min}) \exp\left(-\frac{\tilde{\alpha}(0.1\zeta\lambda^-)^2}{8r^2\gamma_*^4}\right) \quad (15)$$

The second claim follows using $\lambda_k^- \geq \lambda^-$, $f := \lambda^+/\lambda^-$, $\tilde{h}_k := \lambda_{k+1}^+/\lambda_k^-$ in the above expression and applying Lemma 1.4.

Consider the third claim. Using the expression for $\tilde{\mathcal{H}}_k$ given in Definition 5.6, it is easy to see that

$$\|\tilde{\mathcal{H}}_k\|_2 \leq \max\{\|\tilde{H}_k\|_2, \|\tilde{H}_{k,\perp}\|_2\} + \|\tilde{B}_k\|_2 \leq \frac{1}{\alpha} \sum_t e_t e_t' \|_2 + \max(\|T2\|_2, \|T4\|_2) + \|\tilde{B}_k\|_2 \quad (16)$$

where $T2 := \frac{1}{\alpha} \sum_t E_k' \Psi_{k-1} (L_t e_t' + e_t L_t') \Psi_{k-1} E_k$ and $T4 := \frac{1}{\alpha} \sum_t E_{k,\perp}' \Psi_{k-1} (L_t e_t' + e_t L_t') \Psi_{k-1} E_{k,\perp}$. The second inequality follows by using the facts that (i) $\tilde{H}_k = T1 - T2$ where $T1 := \frac{1}{\alpha} \sum_t E_k' \Psi_{k-1} e_t e_t' \Psi_{k-1} E_k$, (ii) $\tilde{H}_{k,\perp} = T3 - T4$ where $T3 := \frac{1}{\alpha} \sum_t E_{k,\perp}' \Psi_{k-1} e_t e_t' \Psi_{k-1} E_{k,\perp}$, and (iii) $\max(\|T1\|_2, \|T3\|_2) \leq \|\frac{1}{\alpha} \sum_t e_t e_t'\|_2$.

Next, we obtain high probability bounds on each of the terms on the RHS of (16) using the Hoeffding corollaries.

Consider $\|\frac{1}{\alpha} \sum_t e_t e_t'\|_2$. Let $Z_t = e_t e_t'$. (a) As explained in the beginning of the proof, conditioned on $X_{j,K,k-1}$, the various Z_t 's in the summation are independent whenever $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$. (b) Conditioned on $X_{j,K,k-1}$, $0 \preceq Z_t \preceq b_1 I$ w.p. one for all $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$. Here $b_1 := \phi^{+2}\zeta$. (c) Using $\|\Phi_K P_j\|_2 \leq (r + c)\zeta$, $0 \preceq \frac{1}{\alpha} \sum_t \mathbf{E}(Z_t | X_{j,K,k-1}) \preceq b_2 I$, $b_2 := (r + c)^2 \zeta^2 \phi^{+2} \lambda^+$ for all $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$.

Thus, applying Corollary 1.6 with $\epsilon = 0.1\zeta\lambda^-$,

$$\mathbf{P}(\|\frac{1}{\alpha} \sum_t e_t e_t'\|_2 \leq b_2 + 0.1\zeta\lambda^- | X_{j,K,k-1}) \geq 1 - n \exp\left(-\frac{\tilde{\alpha}(0.1\zeta\lambda^-)^2}{8 \cdot b_1^2}\right) \text{ for all } X_{j,K,k-1} \in \Gamma_{j,K,k-1} \quad (17)$$

Consider $T2$. Let $Z_t := E_k' \Psi_{k-1} (L_t e_t' + e_t L_t') \Psi_{k-1} E_k$ which is of size $\tilde{c}_k \times \tilde{c}_k$. Then $T2 = \frac{1}{\alpha} \sum_t Z_t$. (a) Conditioned on $X_{j,K,k-1}$, the various Z_t 's used in the summation are mutually independent whenever $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$. (b) Notice that $E_k' \Psi_{k-1} L_t = R_k a_{t,k} + E_k' (D_{\det,k} a_{t,\det} + D_{\text{undet},k} a_{t,\text{undet}})$ and $E_k' \Psi_{k-1} e_t = (R_k^{-1})' D_k' e_t = (R_k^{-1})' D_k' I_{T_t} [(\Phi_K)'_{T_t} (\Phi_K)_{T_t}]^{-1} I_{T_t}' \Phi_K P_j a_t$. Thus conditioned on $X_{j,K,k-1}$, $\|Z_t\|_2 \leq 2b_3$ w.p. one for all $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$. Here, $b_3 := \frac{\sqrt{r\zeta}}{\sqrt{1-r^2\zeta^2}} \phi^+ \gamma_*$. This follows using $\|(R_k^{-1})'\|_2 \leq 1/\sqrt{1-r^2\zeta^2}$, $\|e_t\|_2 \leq \phi^+ \sqrt{\zeta}$ and $\|E_k' \Psi_{k-1} L_t\|_2 \leq \|L_t\|_2 \leq \sqrt{r}\gamma_*$. (c) Also, $\|\frac{1}{\alpha} \sum_t \mathbf{E}(Z_t | X_{j,K,k-1})\|_2 \leq 2b_4$ where $b_4 := \kappa_{s,e}(r + c)\zeta\phi^+(\lambda_k^+ + r\zeta\lambda^+ + \frac{r^2\zeta^2}{\sqrt{1-r^2\zeta^2}}\lambda_{k+1}^+)$.

Thus, applying Corollary 1.7 with $\epsilon = 0.1\zeta\lambda^-$, for all $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$,

$$\mathbf{P}(\|T2\|_2 \leq 2b_4 + 0.1\zeta\lambda^- | X_{j,K,k-1}) \geq 1 - \tilde{c}_k \exp\left(-\frac{\tilde{\alpha}(0.1\zeta\lambda^-)^2}{32 \cdot 4b_3^2}\right)$$

Consider $T4$. Let $Z_t := E_{k,\perp}' \Psi_{k-1} (L_t e_t' + e_t L_t') \Psi_{k-1} E_{k,\perp}$ which is of size $(n - \tilde{c}_k) \times (n - \tilde{c}_k)$. Then $T4 = \frac{1}{\alpha} \sum_t Z_t$. (a) conditioned on $X_{j,K,k-1}$, the various Z_t 's used in the summation are mutually independent whenever $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$. (b) Notice that $E_{k,\perp}' \Psi_{k-1} L_t = E_{k,\perp}' (D_{\det,k} a_{t,\det} + D_{\text{undet},k} a_{t,\text{undet}})$. Thus, conditioned on $X_{j,K,k-1}$, $\|Z_t\|_2 \leq 2b_5$ w.p. one

for all $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$. Here $b_5 := \sqrt{r\zeta}\phi^+\gamma_*$. (c) Also, for all $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$, $\|\frac{1}{\alpha}\sum_t \mathbf{E}(Z_t|X_{j,K,k-1})\|_2 \leq 2b_6$, $b_6 := \kappa_{s,e}(r+c)\zeta\phi^+(\lambda_{k+1}^+ + r\zeta\lambda^+)$. Applying Corollary 1.7 with $\epsilon = 0.1\zeta\lambda^-$, for all $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$,

$$\mathbf{P}(\|T4\|_2 \leq 2b_6 + 0.1\zeta\lambda^-|X_{j,K,k-1}) \geq 1 - (n - \tilde{c}_k) \exp(-\frac{\tilde{\alpha}(0.1\zeta\lambda^-)^2}{32 \cdot 4b_6^2}) \geq 1 - (n - \tilde{c}_{\min}) \exp(-\frac{\tilde{\alpha}(0.1\zeta\lambda^-)^2}{32 \cdot 4b_6^2})$$

Consider $\max(\|T2\|_2, \|T4\|_2)$. Since $b_3 = b_5$ and $b_4 > b_6$, so $2b_6 + \epsilon < 2b_4 + \epsilon$. Therefore, for all $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$,

$$\mathbf{P}(\|T4\|_2 \leq 2b_4 + 0.1\zeta\lambda^-|X_{j,K,k-1}) \geq 1 - (n - \tilde{c}_k) \exp(-\frac{\tilde{\alpha}(0.1\zeta\lambda^-)^2}{32 \cdot 4b_3^2})$$

By union bound, for all $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$,

$$\mathbf{P}(\max(\|T2\|_2, \|T4\|_2) \leq 2b_4 + 0.1\zeta\lambda^-|X_{j,K,k-1}) \geq 1 - n \exp(-\frac{\tilde{\alpha}(0.1\zeta\lambda^-)^2}{32 \cdot 4b_3^2}) \quad (18)$$

Notice that if we also introduce an extra denseness coefficient $\kappa_{s,D} := \max_j \max_k \kappa_s(D_k)$, then $\mathbf{P}(\|T2\|_2 \leq 2\kappa_{s,D}b_4 + 0.1\zeta\lambda^-|X_{j,K,k-1}) \geq 1 - \tilde{c}_k \exp(-\frac{\tilde{\alpha}(0.1\zeta\lambda^-)^2}{32 \cdot 4b_3^2})$. Thus, $\mathbf{P}(\max(\|T2\|_2, \|T4\|_2) \leq 2\max(\kappa_{s,D}b_4, b_6) + 0.1\zeta\lambda^-|X_{j,K,k-1}) \geq 1 - n \exp(-\frac{\tilde{\alpha}(0.1\zeta\lambda^-)^2}{32 \cdot 4b_3^2})$. This would help to get a looser bounds on \tilde{g}_{\max} and \tilde{h}_{\max} in Theorem 4.1.

Consider $\|\tilde{B}_k\|_2$. Let $Z_t := E_{k,\perp}'\Psi_{k-1}(L_t - e_t)(L_t' - e_t')\Psi_{k-1}E_k$ which is of size $(n - \tilde{c}_k) \times \tilde{c}_k$. Then $\tilde{B}_k = \frac{1}{\alpha}\sum_t Z_t$. (a) conditioned on $X_{j,K,k-1}$, the various Z_t 's used in the summation are mutually independent whenever $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$. (b) Notice that $E_{k,\perp}'\Psi_{k-1}(L_t - e_t) = E_{k,\perp}'(D_{\det,k}a_{t,\det} + D_{\text{undet},k}a_{t,\text{undet}} - \Psi_{k-1}e_t)$ and $E_k'\Psi_{k-1}(L_t - e_t) = R_k a_{t,k} + E_k'(D_{\det,k}a_{t,\det} + D_{\text{undet},k}a_{t,\text{undet}} - \Psi_{k-1}e_t)$. Thus, conditioned on $X_{j,K,k-1}$, $\|Z_t\|_2 \leq b_7$ w.p. one for all $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$. Here $b_7 := (\sqrt{r}\gamma_* + \phi^+\sqrt{\zeta})^2$. (c) $\|\frac{1}{\alpha}\sum_t \mathbf{E}(Z_t|X_{j,K,k-1})\|_2 \leq b_8$ for all $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$ where

$$b_8 := (r+c)\zeta\kappa_{s,e}\phi^+\lambda_k^+ + [(r+c)\zeta\kappa_{s,e}\phi^+ + (r+c)\zeta\kappa_{s,e}\frac{r^2\zeta^2}{\sqrt{1-r^2\zeta^2}}]\lambda_{k+1}^+[r^2\zeta^2 + 2(r+c)r\zeta^2\kappa_{s,e}\phi^+ + (r+c)^2\zeta^2\kappa_{s,e}^2\phi^{+2}]\lambda^+$$

Thus, applying Corollary 1.7 with $\epsilon = 0.1\zeta\lambda^-$,

$$\mathbf{P}(\|\tilde{B}_k\|_2 \leq b_8 + 0.1\zeta\lambda^-|X_{j,K,k-1}) \geq 1 - n \exp(-\frac{\tilde{\alpha}(0.1\zeta\lambda^-)^2}{32 \cdot b_7^2}) \text{ for all } X_{j,K,k-1} \in \Gamma_{j,K,k-1} \quad (19)$$

Using (16), (17), (18) and (19) and the union bound, for any $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$,

$$\mathbf{P}(\|\tilde{\mathcal{H}}_k\|_2 \leq b_9 + 0.2\zeta\lambda^-|X_{j,K,k-1}) \geq 1 - \tilde{p}_3(\tilde{\alpha}, \zeta)$$

where $b_9 := b_2 + 2b_4 + b_8$ and

$$\tilde{p}_3(\tilde{\alpha}, \zeta) := n \exp(-\frac{\tilde{\alpha}\epsilon^2}{8 \cdot b_1^2}) + n \exp(-\frac{\tilde{\alpha}\epsilon^2}{32 \cdot 4b_3^2}) + n \exp(-\frac{\tilde{\alpha}\epsilon^2}{32 \cdot b_7^2}) \quad (20)$$

with $b_1 = \phi^{+2}\zeta$, $b_3 := \sqrt{r\zeta}\phi^+\gamma_*$, $b_7 := (\sqrt{r}\gamma_* + \phi^+\sqrt{\zeta})^2$. Using $\lambda_k^- \geq \lambda^-$, $f := \lambda^+/\lambda^-$, $\tilde{g}_k := \lambda_k^+/\lambda_k^-$ and $\tilde{h}_k := \lambda_{k+1}^+/\lambda_k^-$, and then applying Lemma 1.4, the third claim of the lemma follows. ■

VIII. SIMULATION EXPERIMENTS

1) *Data Generation:* The simulated data is generated as follows. The measurement matrix $\mathcal{M}_t := [M_1, M_2, \dots, M_t]$ is of size 2048×5200 . It can be decomposed as a sparse matrix $\mathcal{S}_t := [S_1, S_2, \dots, S_t]$ plus a low rank matrix $\mathcal{L}_t := [L_1, L_2, \dots, L_t]$.

The sparse matrix $\mathcal{S}_t := [S_1, S_2, \dots, S_t]$ is generated as follows. For $1 \leq t \leq t_{\text{train}} = 200$, $S_t = 0$. For $t_{\text{train}} < t \leq 5200$, S_t has s nonzero elements. The initial support $T_0 = \{1, 2, \dots, s\}$. Every Δ time instants we increment the support indices by 1. For example, for $t \in [t_{\text{train}} + 1, t_{\text{train}} + \Delta - 1]$, $T_t = T_0$, for $t \in [t_{\text{train}} + \Delta, t_{\text{train}} + 2\Delta - 1]$, $T_t = \{2, 3, \dots, s+1\}$ and so on. Thus, the support set changes in a highly correlated fashion over time and this results in the matrix \mathcal{S}_t being low rank. The larger the value of Δ , the smaller will be the rank of \mathcal{S}_t (for $t > t_{\text{train}} + \Delta$). The signs of the nonzero elements of S_t are ± 1 with equal probability and the magnitudes are uniformly distributed between 2 and 3. Thus, $S_{\min} = 2$.

The low rank matrix $\mathcal{L}_t := [L_1, L_2, \dots, L_t]$ where $L_t := P_{(t)}a_t$ is generated as follows: There are a total of $J = 2$ subspace change times, $t_1 = 301$ and $t_2 = 2501$. $r_0 = 36$, $c_{1,\text{new}} = c_{2,\text{new}} = 1$ and $c_{1,\text{old}} = c_{2,\text{old}} = 3$. Let U be an $2048 \times (r_0 + c_{1,\text{new}} + c_{2,\text{new}})$ orthonormalized random Gaussian matrix. For $1 \leq t \leq t_1 - 1$, $P_{(t)} = P_0$ has rank r_0 with

$P_0 = U_{[1,2,\dots,36]}$. For $t_1 \leq t \leq t_2 - 1$, $P_{(t)} = P_1 = [P_0 \setminus P_{1,\text{old}} \ P_{1,\text{new}}]$ has rank $r_1 = r_0 + c_{1,\text{new}} - c_{1,\text{old}} = 34$ with $P_{1,\text{new}} = U_{[37]}$ and $P_{1,\text{old}} = U_{[9,18,36]}$. For $t \geq t_2$, $P_{(t)} = P_2 = [P_1 \setminus P_{2,\text{old}} \ P_{2,\text{new}}]$ has rank $r_2 = r_1 + c_{2,\text{new}} - c_{2,\text{old}} = 32$ with $P_{2,\text{new}} = U_{[38]}$ and $P_{2,\text{old}} = U_{[8,17,35]}$. a_t is independent over t . The various $(a_t)_i$'s are also mutually independent for different i . For $1 \leq t < t_1$, we let $(a_t)_i$ be uniformly distributed between $-\gamma_{i,t}$ and $\gamma_{i,t}$, where

$$\gamma_{i,t} = \begin{cases} 400 & \text{if } i = 1, 2, \dots, 9, \forall t, \\ 30 & \text{if } i = 10, 11, \dots, 18, \forall t. \\ 2 & \text{if } i = 19, 20, \dots, 27, \forall t. \\ 1 & \text{if } i = 28, 29 \dots, 36, \forall t. \end{cases} \quad (21)$$

For $t_1 \leq t < t_2$, $a_{t,*}$ is an $r_0 - c_{1,\text{old}}$ length vector, $a_{t,\text{new}}$ is a $c_{1,\text{new}}$ length vector and $L_t := P_{(t)}a_t = P_1a_t = (P_0 \setminus P_{1,\text{old}})a_{t,*} + P_{1,\text{new}}a_{t,\text{new}}$. Now, $(a_{t,*})_i$ is uniformly distributed between $-\gamma_{i,t}$ and $\gamma_{i,t}$ for $i = 1, 2, \dots, 35$ and $a_{t,\text{new}}$ is uniformly distributed between $-\gamma_{\text{new},t}$ and $\gamma_{\text{new},t}$, where

$$\gamma_{i,t} = \begin{cases} 400 & \text{if } i = 1, 2, \dots, 8, \forall t, \\ 30 & \text{if } i = 9, 10, \dots, 16, \forall t. \\ 2 & \text{if } i = 17, 18, \dots, 24, \forall t. \\ 1 & \text{if } i = 25, 26, \dots, 33, \forall t. \end{cases}$$

$$\gamma_{\text{new},t} = \begin{cases} 1.1^{k-1} & \text{if } t_1 + (k-1)\alpha \leq t \leq t_1 + k\alpha - 1, k = 1, 2, 3, 4, \\ 1.1^{4-1} = 1.331 & \text{if } t \geq t_1 + 4\alpha. \end{cases} \quad (22)$$

For $t \geq t_2$, $a_{t,*}$ is an $r_1 - c_{2,\text{old}}$ length vector, $a_{t,\text{new}}$ is a $c_{2,\text{new}}$ length vector and $L_t := P_{(t)}a_t = P_2a_t = [P_0 \setminus P_{1,\text{old}} \ P_{1,\text{new}}]a_{t,*} + P_{2,\text{new}}a_{t,\text{new}}$. Also, $(a_{t,*})_i$ is uniformly distributed between $-\gamma_{i,t}$ and $\gamma_{i,t}$ for $i = 1, 2, \dots, r_1 - c_{2,\text{old}}$ and $a_{t,\text{new}}$ is uniformly distributed between $-\gamma_{\text{new},t}$ and $\gamma_{\text{new},t}$ where

$$\gamma_{i,t} = \begin{cases} 400 & \text{if } i = 1, 2, \dots, 7, \forall t, \\ 30 & \text{if } i = 8, 9, \dots, 14, \forall t. \\ 2 & \text{if } i = 15, 16, \dots, 21, \forall t. \\ 1.331 & \text{if } i = 22, \forall t. \\ 1 & \text{if } i = 23, 24, \dots, 31, \forall t. \end{cases} \quad (23)$$

$$\gamma_{\text{new},t} = \begin{cases} 1.1^{k-1} & \text{if } t_2 + (k-1)\alpha \leq t \leq t_2 + k\alpha - 1, k = 1, 2, \dots, 7, \\ 1.1^{7-1} = 1.7716 & \text{if } t \geq t_2 + 7\alpha. \end{cases} \quad (24)$$

Thus for the above model, $S_{\min} = 2$, $\gamma_* = 400$, $\gamma_{\text{new}} = 1$, $\lambda^+ = 53333$, $\lambda^- = 0.3333$ and $f := \frac{\lambda^+}{\lambda^-} = 1.6 \times 10^5$. One way to get the clusters of $\{1, 2, \dots, r_j\}$ is as follows.

- 1) For $t_1 \leq t < t_2$ with $j = 1$, let $\mathcal{G}_{1,(1)} = \{1, 2, \dots, 8\}$, $\mathcal{G}_{1,(2)} = \{9, 10, \dots, 16\}$ and $\mathcal{G}_{1,(3)} = \{17, 18, \dots, 34\}$. Thus, $\tilde{c}_{1,1} = \tilde{c}_{1,2} = 8$, $\tilde{c}_{1,3} = 18$, $\tilde{g}_{j,1} = \tilde{g}_{j,2} = 1$, $\tilde{g}_{j,3} = 4$, $\tilde{h}_{j,1} = 0.0056$, $\tilde{h}_{j,2} = 0.0044$.
- 2) For $t \geq t_2$ with $j = 2$, let $\mathcal{G}_{1,(1)} = \{1, 2, \dots, 7\}$, $\mathcal{G}_{1,(2)} = \{8, 10, \dots, 14\}$ and $\mathcal{G}_{1,(3)} = \{17, 18, \dots, 32\}$. Thus, $\tilde{c}_{1,1} = \tilde{c}_{1,2} = 7$, $\tilde{c}_{1,3} = 16$, $\tilde{g}_{j,1} = \tilde{g}_{j,2} = 1$, $\tilde{g}_{j,3} = 4$, $\tilde{h}_{j,1} = 0.0056$, $\tilde{h}_{j,2} = 0.0044$.
- 3) Therefore, $\tilde{g}_{\max} = 4$, $\tilde{h}_{\max} = 0.0056$ and $\tilde{c}_{\min} = 7$.

We used $\mathcal{L}_{t_{\text{train}}} + \mathcal{N}_{t_{\text{train}}}$ as the training sequence to estimate \hat{P}_0 . Here $\mathcal{N}_{t_{\text{train}}} = [N_1, N_2, \dots, N_{t_{\text{train}}}]$ is i.i.d. random noise with each $(N_t)_i$ uniformly distributed between -10^{-3} and 10^{-3} . This is done to ensure that $\text{span}(\hat{P}_0) \neq \text{span}(P_0)$ but only approximates it.

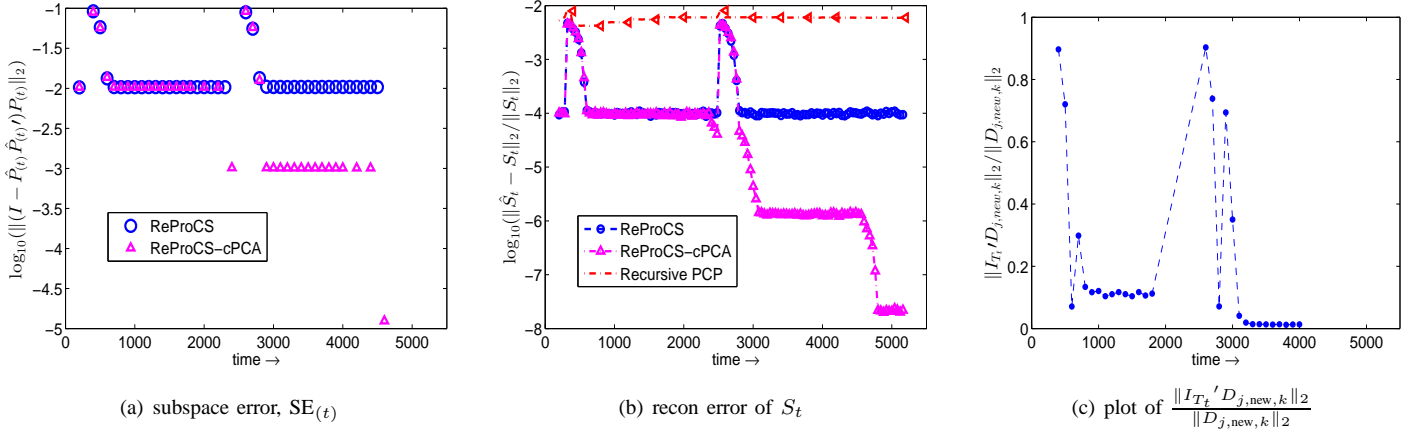


Fig. 4. $r_0 = 36$, $s = \max_t |T_t| = 20$ and $\Delta = 10$. The times at which PCP is done are marked by red triangles in (b).

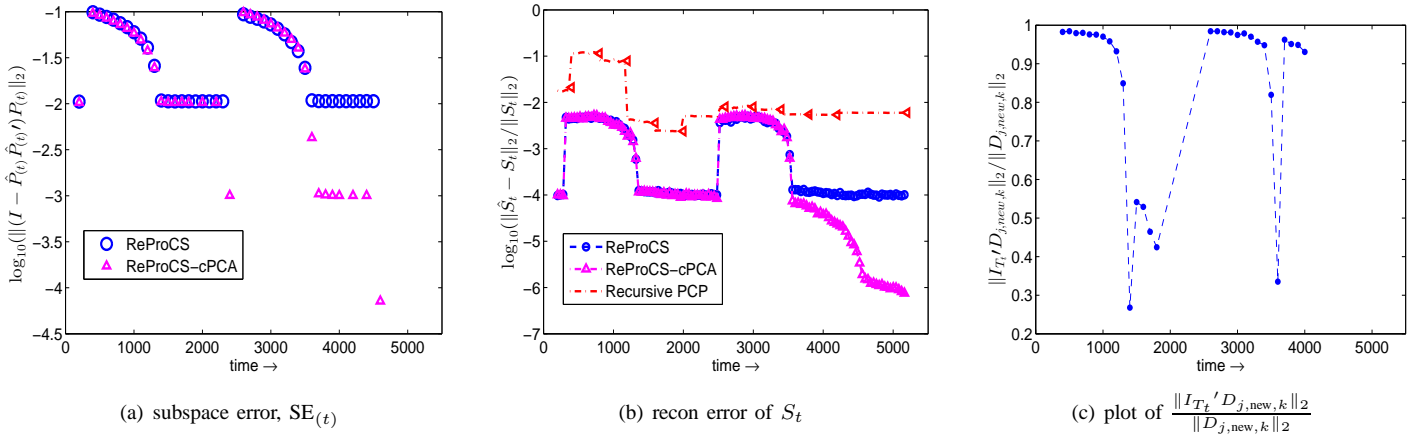


Fig. 5. $r_0 = 36$, $s = \max_t |T_t| = 20$ and $\Delta = 50$. The times at which PCP is done are marked by red triangles in (b).

2) *Results:* For Fig. 4 and Fig. 5, we used $s = 20$. We used $\Delta = 10$ for Fig. 4 and $\Delta = 50$ for Fig. 5. Because of the correlated support change, the $2048 \times t$ sparse matrix $\mathcal{S}_t = [S_1, S_2, \dots, S_t]$ is rank deficient in either case, e.g. for Fig. 4, \mathcal{S}_t has rank 29, 39, 49, 259 at $t = 300, 400, 500, 2600$; for Fig. 5, \mathcal{S}_t has rank 21, 23, 25, 67 at $t = 300, 400, 500, 2600$. We plot the subspace error $SE(t)$ and the normalized error for S_t , $\frac{\|\hat{S}_t - S_t\|_2}{\|S_t\|_2}$ averaged over 100 Monte Carlo simulations.

As can be seen from Fig. 4 and Fig. 5, the subspace error $SE(t)$ of ReProCS and ReProCS-cPCA decreased exponentially and stabilized. Furthermore, ReProCS-cPCA outperforms over ReProCS greatly when deletion steps are done (i.e., at $t = 2400$ and 4600). The averaged normalized error for S_t followed a similar trend.

We also compared against PCP [2]. At every $t = t_j + 4k\alpha$, we solved (1) with $\lambda = 1/\sqrt{\max(n, t)}$ as suggested in [2] to recover \mathcal{S}_t and \mathcal{L}_t . We used the estimates of S_t for the last 4α frames as the final estimates of \hat{S}_t . So, the \hat{S}_t for $t = t_j + 1, \dots, t_j + 4\alpha$ is obtained from PCP done at $t = t_j + 4\alpha$, the \hat{S}_t for $t = t_j + 4\alpha + 1, \dots, t_j + 8\alpha$ is obtained from PCP done at $t = t_j + 8\alpha$ and so on. Because of the correlated support change, the error of PCP was larger in both cases.

We also plot the ratio $\frac{\|I_{T_t}' D_{j,new,k}\|_2}{\|D_{j,new,k}\|_2}$ at the projection PCA times. This serves as a proxy for $\kappa_s(D_{j,new,k})$ (which has exponential computational complexity). As can be seen from Fig. 4 and Fig. 5, this ratio is less than 1 and it becomes larger when Δ increases (T_t becomes more correlated over t).

We implemented ReProCS-cPCA using Algorithm 2 with $\alpha = 100$, $\tilde{\alpha} = 200$ and $K = 15$. The algorithm is not very sensitive to these choices. Also, we let $\xi = \xi_t$ and $\omega = \omega_t$ vary with time. Recall that ξ_t is the upper bound on $\|\beta_t\|_2$. We do not know β_t . All we have is an estimate of β_t from $t-1$, $\hat{\beta}_{t-1} = (I - \hat{P}_{t-1} \hat{P}_{t-1}') \hat{L}_{t-1}$. We used a value a little larger than $\|\hat{\beta}_{t-1}\|_2$; we let $\xi_t = 2\|\hat{\beta}_{t-1}\|_2$. The parameter ω_t is the support estimation threshold. One reasonable way to pick this is to use a percentage energy threshold of $\hat{S}_{t,cs}$ [37]. For a vector v , define the 99%-energy set of v as $T_{0.99}(v) := \{i : |v_i| \geq v^{0.99}\}$ where the 99%

energy threshold, $v^{0.99}$, is the largest value of $|v_i|$ so that $\|v_{T_{0.99}}\|_2^2 \geq 0.99\|v\|_2^2$. It is computed by sorting $|v_i|$ in non-increasing order of magnitude. One keeps adding elements to $T_{0.99}$ until $\|v_{T_{0.99}}\|_2^2 \geq 0.99\|v\|_2^2$. We used $\omega_t = 0.5(\hat{S}_{t,cs})^{0.99}$.

IX. CONCLUSIONS AND FUTURE WORK

We studied the problem of recursive sparse recovery in the presence of large but structured noise (noise lying in a “slowly changing” low dimensional subspace). We introduced the ReProCS with cluster-PCA (ReProCS-cPCA) algorithm that addresses some of the limitations of our earlier work on ReProCS [21] and of PCP [2]. Under mild assumptions, we showed that, w.h.p., ReProCS-cPCA can exactly recover the support set of S_t at all times; and the reconstruction errors of both S_t and L_t are upper bounded by a time-invariant and small value at all times. In ongoing work, we are studying the undersampled measurements case. Open questions include (i) how to analyze a practical version of ReProCS-cPCA (which does not assume knowledge of signal model parameters), and (ii) how to study the correlated a_t ’s case (e.g. the case where a_t ’s satisfy a linear random walk model). The starting point for (ii) would be to try to use the matrix Azuma inequality [26] instead of Hoeffding.

APPENDIX A

PROOF OF LEMMA 6.1

The proof follows by using the following three lemmas.

Lemma A.1 (Exponential decay of ζ_k^+): Assume that all the conditions of Theorem 4.1 hold. Let $\zeta_*^+ = r\zeta$. Define the series ζ_k^+ as in Definition 5.3. Then,

- 1) $\zeta_0^+ = 1$ and $\zeta_k^+ \leq 0.6^k + 0.4c\zeta$ for all $k = 1, 2, \dots, K$,
- 2) the denominator of ζ_k^+ is positive for all $k = 1, 2, \dots, K$.

Proof: This lemma is the same as [21, Lemma 37] but with ζ_*^+ defined differently. ■

Lemma A.2 (Sparse recovery, support recovery and expression for e_t): Assume that all conditions of Theorem 4.1 hold.

- 1) If $\zeta_* \leq \zeta_*^+ := r\zeta$ and $\zeta_{k-1} \leq \zeta_{k-1}^+ \leq 0.6^{k-1} + 0.4c\zeta$, then for all $t \in \mathcal{I}_{j,k}$, for any $k = 1, 2, \dots, K$,
 - a) the projection noise β_t satisfies $\|\beta_t\|_2 \leq \zeta_{k-1}^+ \sqrt{c}\gamma_{\text{new},k} + \zeta_*^+ \sqrt{r}\gamma_* \leq \sqrt{c}0.72^{k-1}\gamma_{\text{new}} + 1.06\sqrt{\zeta} \leq \xi$.
 - b) the CS error satisfies $\|\hat{S}_{t,cs} - S_t\|_2 \leq 7\xi$.
 - c) $\hat{T}_t = T_t$
 - d) e_t satisfies (8) and $\|e_t\|_2 \leq \phi^+[\kappa_s^+ \zeta_{k-1}^+ \sqrt{c}\gamma_{\text{new},k} + \zeta_*^+ \sqrt{r}\gamma_*] \leq 0.18 \cdot 0.72^{k-1} \sqrt{c}\gamma_{\text{new}} + 1.17 \cdot 1.06\sqrt{\zeta}$
- 2) For all $k = 1, 2, \dots, K$, $\mathbf{P}(\hat{T}_t = T_t \text{ and } e_t \text{ satisfies (8) for all } t \in \mathcal{I}_{j,k} | X_{j,k-1,0}) = 1$ for all $X_{j,k-1,0} \in \Gamma_{j,k-1,0}$.
- 3) For all $k = 1, 2, \dots, K$, $\mathbf{P}(\hat{T}_t = T_t \text{ and } e_t \text{ satisfies (8) for all } t \in \mathcal{I}_{j,k} | \Gamma_{j,k-1,0}^e) = 1$.

Proof: The first claim is the same as [21, Lemma 30] but with ζ_*^+ defined differently. The proof follows in an analogous fashion. The second claim follows from the first using Remark 6.3. The third claim follows using Lemma 1.4. ■

Lemma A.3 (High probability bound on ζ_k): Assume that all the conditions of Theorem 4.1 hold. Let $\zeta_*^+ = r\zeta$. Then, for all $k = 1, 2, \dots, K$,

$$\mathbf{P}(\zeta_k \leq \zeta_k^+ | \Gamma_{j,k-1,0}^e) \geq p_k(\alpha, \zeta)$$

where ζ_k^+ is defined in Definition 5.3 and $p_k(\alpha, \zeta)$ is defined in [21, Lemma 35].

Proof: Using Lemma A.1, (i) $\zeta_0^+ = 1$ and $\zeta_{k-1}^+ \leq 0.6^{k-1} + 0.4c\zeta$ and (ii) the denominator of ζ_k^+ is positive. Using this and the theorem’s conditions, the above lemma follows exactly as in [21, Lemma 35]. The only difference is that ζ_*^+ is defined differently. Also, $\Gamma_{j,k} := \Gamma_{j,k,0}$. The proof proceeds by first bounding ζ_k (in a fashion similar to the bound in Lemma 7.6); using Lemma A.2 to get an expression for e_t ; and finally using Corollaries 1.6 and 1.7 to get high probability bounds on each of the terms in the bound on ζ_k . ■

Proof of Lemma 6.1: Lemma 6.1 follows by combining Lemma A.3 and the third claim of Lemma A.2 and using the fact that $\mathbf{P}(\Gamma_{j,k,0}^e | \Gamma_{j,k-1,0}^e) = \mathbf{P}(\zeta_k \leq \zeta_k^+, \hat{T}_t = T_t \text{ and } e_t \text{ satisfies (8) for all } t \in \mathcal{I}_{j,k} | \Gamma_{j,k-1,0}^e)$. ■

APPENDIX B

PROOF OF LEMMA 7.3

Proof of Lemma 7.3:

- 1) The first claim follows because $\|D_{\det,k}\|_2 = \|\Psi_{k-1}G_{\det,k}\|_2 = \|\Psi_{k-1}[G_1G_2\cdots G_{k-1}]\|_2 \leq \sum_{k_1=1}^{k-1} \|\Psi_{k-1}G_{k_1}\|_2 \leq \sum_{k_1=1}^{k-1} \|\Psi_{k_1}G_{k_1}\|_2 = \sum_{k_1=1}^{k-1} \tilde{\zeta}_{k_1} \leq \sum_{k_1=1}^{k-1} \tilde{c}_{k_1}\zeta \leq r\zeta$. The first inequality follows by triangle inequality. The second one follows because $\hat{G}_1, \dots, \hat{G}_{k-1}$ are mutually orthonormal and so $\Psi_{k-1} = \prod_{k_2=1}^{k-1} (I - \hat{G}_{k_2}\hat{G}_{k_2}')$.
- 2) By the first claim, $\|(I - \hat{G}_{\det,k}\hat{G}_{\det,k}')G_{\det,k}\|_2 = \|\Psi_{k-1}G_{\det,k}\|_2 \leq r\zeta$. By item 2) of Lemma 1.12 with $P = G_{\det,k}$ and $\hat{P} = \hat{G}_{\det,k}$, the result $\|G_{\det,k}G_{\det,k}' - \hat{G}_{\det,k}\hat{G}_{\det,k}'\|_2 \leq 2r\zeta$ follows.
- 3) Recall that $D_k \stackrel{QR}{=} E_k R_k$ is a QR decomposition where E_k is orthonormal and R_k is upper triangular. Therefore, $\sigma_i(D_k) = \sigma_i(R_k)$. Since $\|(I - \hat{G}_{\det,k}\hat{G}_{\det,k}')G_{\det,k}\|_2 = \|\Psi_{k-1}G_{\det,k}\|_2 \leq r\zeta$ and $G_k'G_{\det,k} = 0$, by item 4) of Lemma 1.12 with $P = G_{\det,k}$, $\hat{P} = \hat{G}_{\det,k}$ and $Q = G_k$, we have $\sqrt{1 - r^2\zeta^2} \leq \sigma_i((I - \hat{G}_{\det,k}\hat{G}_{\det,k}')G_k) = \sigma_i(D_k) \leq 1$.
- 4) Since $D_k \stackrel{QR}{=} E_k R_k$, so $\|D_{\det,k}'E_k\|_2 = \|D_{\det,k}'D_k R_k^{-1}\|_2 = \|G_{\det,k}'\Psi_{k-1}G_k R_k^{-1}\|_2 = \|G_{\det,k}'\Psi_{k-1}G_k R_k^{-1}\|_2 = \|G_{\det,k}'\Psi_{k-1}G_k R_k^{-1}\|_2 = \|G_{\det,k}'D_k R_k^{-1}\|_2 = \|G_{\det,k}'E_k\|_2$. Since $E_k = D_k R_k^{-1} = (I - \hat{G}_{\det,k}\hat{G}_{\det,k}')G_k R_k^{-1}$,

$$\begin{aligned} \|G_{\det,k}'E_k\|_2 &= \|G_{\det,k}'(I - \hat{G}_{\det,k}\hat{G}_{\det,k}')G_k R_k^{-1}\|_2 \\ &\leq \|G_{\det,k}'(I - \hat{G}_{\det,k}\hat{G}_{\det,k}')G_k\|_2 (1/\sqrt{1 - r^2\zeta^2}) = \|G_{\det,k}'\hat{G}_{\det,k}\hat{G}_{\det,k}'G_k\|_2 (1/\sqrt{1 - r^2\zeta^2}) \end{aligned}$$

By item 3) of Lemma 1.12 with $P = G_{\det,k}$, $\hat{P} = \hat{G}_{\det,k}$ and $Q = G_{\det,k}$, we get $\|G_{\det,k}'\hat{G}_{\det,k}\|_2 \leq r\zeta$. By item 3) of Lemma 1.12 with $\hat{P} = \hat{G}_{\det,k}$ and $Q = G_k$, we get $\|\hat{G}_{\det,k}'G_k\|_2 \leq r\zeta$. Therefore, $\|G_{\det,k}'E_k\|_2 = \|E_k'G_{\det,k}\|_2 \leq \frac{r^2\zeta^2}{\sqrt{1 - r^2\zeta^2}}$. ■

REFERENCES

- [1] F. D. L. Torre and M. J. Black, "A framework for robust subspace learning," *International Journal of Computer Vision*, vol. 54, pp. 117–142, 2003.
- [2] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of ACM*, vol. 58, no. 3, 2011.
- [3] J. Wright and Y. Ma, "Dense error correction via l1-minimization," *IEEE Trans. on Info. Th.*, vol. 56, no. 7, pp. 3540–3560, 2010.
- [4] J. Laska, M. Davenport, and R. Baraniuk, "Exact signal recovery from sparsely corrupted measurements through the pursuit of justice," in *Asilomar Conf. on Sig. Sys. Comp.*, Nov 2009, pp. 1556–1560.
- [5] N. H. Nguyen and T. D. Tran, "Robust lasso with missing and grossly corrupted observations," *To appear in IEEE Transaction on Information Theory*, 2012.
- [6] D. Skocaj and A. Leonardis, "Weighted and robust incremental method for subspace learning," in *IEEE Intl. Conf. on Computer Vision (ICCV)*, vol. 2, Oct 2003, pp. 1494–1501.
- [7] Y. Li, L. Xu, J. Morphet, and R. Jacobs, "An integrated algorithm of incremental and robust pca," in *IEEE Intl. Conf. Image Proc. (ICIP)*, 2003, pp. 245–248.
- [8] M. McCoy and J. Tropp, "Two proposals for robust pca using semidefinite programming," *arXiv:1012.1086v3*, 2010.
- [9] H. Xu, C. Caramanis, and S. Sanghavi, "Robust pca via outlier pursuit," *IEEE Tran. on Information Theory*, vol. 58, no. 5, 2012.
- [10] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky, "Rank-sparsity incoherence for matrix decomposition," *SIAM Journal on Optimization*, vol. 21, 2011.
- [11] M. B. McCoy and J. A. Tropp, "Sharp recovery bounds for convex deconvolution, with applications," *arXiv:1205.1580*.
- [12] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Foundations of Computational Mathematics*, no. 6, 2012.
- [13] Y. Hu, S. Goud, and M. Jacob, "A fast majorize-minimize algorithm for the recovery of sparse and low-rank matrices," *IEEE Transactions on Image Processing*, vol. 21, no. 2, p. 742–753, Feb 2012.
- [14] A. E. Waters, A. C. Sankaranarayanan, and R. G. Baraniuk, "Sparcs: Recovering low-rank and sparse matrices from compressive measurements," in *Proc. of Neural Information Processing Systems (NIPS)*, 2011.
- [15] E. Richard, P.-A. Savalle, and N. Vayatis, "Estimation of simultaneously sparse and low rank matrices," *arXiv:1206.6474*, appears in *Proceedings of the 29th International Conference on Machine Learning (ICML 2012)*.
- [16] D. Hsu, S. M. Kakade, and T. Zhang, "Robust matrix decomposition with outliers," *arXiv:1011.1518*.
- [17] M. Mardani, G. Mateos, and G. B. Giannakis, "Recovery of low-rank plus compressed sparse matrices with application to unveiling traffic anomalies," *arXiv:1204.6537*.
- [18] J. Wright, A. Ganesh, K. Min, and Y. Ma, "Compressive principal component pursuit," *arXiv:1202.4596*.
- [19] A. Ganesh, K. Min, J. Wright, and Y. Ma, "Principal component pursuit with reduced linear measurements," *arXiv:1202.6445*.

- [20] M. Tao and X. Yuan, "Recovering low-rank and sparse components of matrices from incomplete and noisy observations," *SIAM Journal on Optimization*, vol. 21, no. 1, pp. 57–81, 2011.
- [21] C. Qiu, N. Vaswani, and L. Hogben, "Recursive robust pca or recursive sparse recovery in large but structured noise," *arXiv: 1211.3754[cs.IT]*, submitted to *IEEE Tran. Info. Th.*, shorter version to appear in *ICASSP 2013*.
- [22] E. Candes, "The restricted isometry property and its implications for compressed sensing," *Compte Rendus de l'Academie des Sciences, Paris, Serie I*, pp. 589–592, 2008.
- [23] T. Zhang and G. Lerman, "A novel m-estimator for robust pca," *arXiv:1112.4863v1*, 2011.
- [24] C. Davis and W. M. Kahan, "The rotation of eigenvectors by a perturbation. iii," *SIAM Journal on Numerical Analysis*, Mar. 1970.
- [25] R. Horn and C. Johnson, *Matrix Analysis*. Cambridge University Press, 1985.
- [26] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Foundations of Computational Mathematics*, vol. 12, no. 4, 2012.
- [27] B. Nadler, "Finite sample approximation results for principal component analysis: A matrix perturbation approach," *The Annals of Statistics*, vol. 36, no. 6, 2008.
- [28] C. Qiu and N. Vaswani, "Real-time robust principal components' pursuit," in *Allerton Conference on Communication, Control, and Computing*, 2010.
- [29] —, "Recursive sparse recovery in large but correlated noise," in *48th Allerton Conference on Communication Control and Computing*, 2011.
- [30] M. Brand, "Incremental singular value decomposition of uncertain data with missing values," in *European Conference on Computer Vision*, 2002, pp. 707–720.
- [31] J. He, L. Balzano, and A. Szlam, "Incremental gradient on the grassmannian for online foreground and background separation in subsampled video," in *IEEE Conf. on Comp. Vis. Pat. Rec. (CVPR)*, 2012.
- [32] E. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. Info. Th.*, vol. 51(12), pp. 4203 – 4215, Dec. 2005.
- [33] Y. Jin and B. Rao, "Algorithms for robust linear regression by exploiting the connection to sparse signal recovery," in *IEEE Intl. Conf. Acoustics, Speech, Sig. Proc. (ICASSP)*, 2010.
- [34] K. Mitra, A. Veeraraghavan, and R. Chellappa, "A robust regression using sparse learning for high dimensional parameter estimation problems," in *IEEE Intl. Conf. Acous. Speech. Sig.Proc.(ICASSP)*, 2010.
- [35] G. Grimmett and D. Stirzaker, *Probability and Random Processes*. Oxford University Press, 2001.
- [36] G. Li and Z. Chen., "Projection-pursuit approach to robust dispersion matrices and principal components: Primary theory and monte carlo," *Journal of the American Statistical Association*, vol. 80, no. 391, pp. 759–766, 1985.
- [37] N. Vaswani and W. Lu, "Modified-cs: Modifying compressive sensing for problems with partially known support," *IEEE Trans. Signal Processing*, September 2010.